University of Alberta

Connectionist Models of Discrimination Learning

by

Leanne Ruth Willson ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment
of the requirements for the degree of *Doctor of Philosophy*

Department of Psychology

Edmonton, Alberta

Fall, 2001

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-69019-9

Canada

University of Alberta

Library Release Form

Name of Author: Leanne R. Willson

Title of Thesis: Connectionist Models of Discrimination Learning

Degree: Doctor of Philosophy

Year this Degree Granted: 2001

Permission is hereby granted to the University of Alberta to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

Leanne R. Willson
9140 - 83 St.
Edmonton, Alberta
T6C 2Z4
Canada

Sept 27, 01

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled Connectionist Models of Discrimination Learning by Leanne Ruth Willson in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Michael R. W. Dawson
Supervisor

Connie K. Varnhagen

C. Donald Heth

David Pierce

E. James Kehoe
External Examiner

26 September, 2001

Abstract

The purpose of this thesis is to explore connectionism in the context of discrimination learning, as a potentially more complete description of associative learning than any of its predecessors, including contingency theory, given that it provides accounts of learning at each of Marr's (1982) three levels of explanation: the implementational level, the algorithmic level, and the computational level. While the computational level explanation of what networks compute is not defined by a specific formula, given that it depends on the particular connectionist network considered, neural networks are certainly not limited to implementational accounts of computational level contingency theory, as has been suggested in the field of associative learning (Shanks, 1995). Given the suggestion that connectionist models can be considered potentially complete accounts of learning, applications of neural networks in learning theory and research are evaluated in this thesis. The particular discrimination learning tasks explored are negative and positive patterning. Two studies regarding the computational power of neural networks with distributed representations are presented, concluding that these models can provide interesting and computationally feasible accounts of negative and positive patterning, without overfitting the data. Six studies are presented concerning the relevance and necessary features of networks prone to catastrophic forgetting. It is concluded that, while catastrophic forgetting may be seen in neural networks in some domains, the phenomenon may not be particularly relevant to the field of discrimination learning when savings are considered through retraining. Finally, four studies are presented in the context of evaluating neural networks through data fitting. A model for evaluation is presented in which a particular connectionist model is evaluated, limitations of that model or of the simulation situation are considered, and another member of the family of connectionist models is considered until an approximate fit to the data is achieved. In this way, one can determine a functionally plausible model for a particular learning situation. It is concluded that connectionism can provide a plausible and powerful account of discrimination learning.

Acknowledgements

I would like to thank my supervisor, Mike Dawson for encouragement and a limitless number of ideas along the way, for knowing when to turn up the pressure and knowing when to dial it back down, and for expecting me to do well. I would also like to thank the members of my supervisory committee: Connie Varnhagen for constant encouragement from the beginning, and for teaching mentorship; and Don Heth for his role in shaping my interest in animal learning and his support from Day 1.

I have benefitted from many supportive relationships, both personally and professionally, through my years of graduate school including those with the members of the Biological Computation Project. Thanks to David Medler and David McCaughan who initiated me into the inner sanctum of the BCP and showed me the ropes. Thanks too to lab-mate Darren Piercey for research discussions that made me think, for input and for coffee. Thanks must go to Walter Bischof who provided much in the way of motivation in the last 6 months of this process. Much thanks for support and great conversations with the valued members of the "Glee Club" through the years – our non-lab-mate member Jan Snyder and lab-mates Patricia Boechler, Jacqueline Leighton, Corinne Zimmerman and especially Monica Valsangkar-Smyth, who was a true support and a wonderful friend from the beginning.

I would like to acknowledge that my research (and I) have been partially supported by an NSERC PGS-B, a Walter Johns Graduate Award, and a University of Alberta Graduate Student Scholarship.

Special thanks go to my family, for whom I am so grateful. Thanks to Leona and Glenn Willoughby, my mom and dad, who have instilled in me a respect for education and achievement, but even more have given support and encouragement to me for this enterprise – and have never once asked what in the world I'm still doing in school. Much thanks too, to Lara Capes, my sister, who provided much needed oases of joy in the midst of the desert of writing this beast. Special thanks to my lovely daughters, Mandy Blue and Emma Rose who have patiently watched this thesis be written (literally, in our dining room) and have enthusiastically cheered each milestone I reached in this process. Finally, I thank my husband, Tim Willson, without whose love and support I surely would not have finished. Thank you for picking up the slack at home, for much proofreading over the years, and for believing in me.

Table of Contents

Chapter 1: INTRODUCTION 1

Chapter 2: THE LEARNING OF ASSOCIATIONS 6

- Behavioural Theories of Associative Learning: Contiguity 7
  - Ebbinghaus 8
  - Pavlov 9
  - Thorndike 11
  - Watson 13
  - Guthrie 15
  - Hull 17
  - Summary 18

- Behavioural Theories of Associative Learning: Contingency 20
  - The Rescorla-Wagner Model 21
  - Contiguity and Contingency Today 23

- Associative Learning: The Cognitive Perspective 23
  - Cognition in Behavioural Approaches 24
  - Modern Cognitive Psychology and Learning 25

- Summary 26

Chapter 3: CONNECTIONISM: LEVELS OF ANALYSIS 28

- Basic Principles of PDP Models 29
  - Processing Units 29
    - Neuronal Inspiration 29
    - Activation Functions 29
  - Network Topology 32
  - Learning in PDP Networks 32

- Sample PDP Networks 34
  - Perceptrons 34
  - Multi-layer Perceptrons 36

- Associative Learning and Connectionist Learning Rules 37

- Connectionism and the Tri-Level Hypothesis 41
  - Connectionism: an Implementational Account 44

Connectionism: an Algorithmic Level Account........ 46
Connectionism: a Computational Level Account........ 47

Connectionism as Cognitive Theory........ 47
The Connectionist Black Box........ 50

General Discussion........ 52

Chapter 4: THE "PROBLEM OF PATTERNING" REVISITED........ 56

Solutions to the Problem of Patterning........ 58
Patterning: A Unique Cue Account........ 58
Patterning: A Configural Cue Account........ 59
Elemental and Configural Processes........ 59

Linear Regression Model........ 61
Ballard's Packing Constraint........ 64
Solving the Packing Problem........ 65
The Inductive Approach........ 67

Study 1: Configural Cues in Patterning........ 69
Method........ 71
Networks........ 71
Input Coding........ 72
Training........ 72
Pretraining........ 73
Patterning........ 73
Results and Discussion........ 73

Study 2: Initial Excitatory Summation........ 76
Method........ 76
Networks........ 76
Input Coding........ 76
Training........ 77
Results........ 78
Discussion........ 79
Interpreting the Network........ 80

General Discussion........ 82

Chapter 5: CATASTROPHIC FORGETTING........ 86

Why do Networks Forget?........ 88
Study 3-A: Forgetting or Savings: Rate of Re-acquisition........ 89

How is Forgetting Measured?.......... 90
Is This Forgetting Reasonable or Catastrophic?.......... 91
Forgetting or Remembering?.......... 92
Method.......... 93
    Networks.......... 93
    Training.......... 93
Results and Discussion.......... 95

Study 3-B: Forgetting or Savings: What Does Remembering Look Like?.. 98
Method.......... 98
Results and Discussion.......... 99

Study 4: Separate Elements, Separate Tasks.......... 102
Method.......... 102
    Networks.......... 102
    Input Coding.......... 102
    Training.......... 103
Results and Discussion.......... 104

Study 5-A: Local and Distributed Representations: Savings.......... 105
Method.......... 106
    Networks.......... 106
    Training.......... 106
Results and Discussion.......... 107

Study 5-B: Local and Distributed Representations: "Goodness".......... 108
Method.......... 109
    Networks.......... 109
    Training.......... 109
Results and Discussion.......... 110

Study 5-C: Local and Distributed Representations: Network Interpretation.......... 111
Method.......... 111
Results and Discussion.......... 112

General Discussion.......... 116

Chapter 6: EVALUATING CONNECTIONIST MODELS OF LEARNING.. 119

Study 6: A Perceptron Solution to Negative Patterning.......... 123
Method.......... 126
    Networks.......... 126
    Input Coding.......... 127
    Training.......... 128

      - Pretraining 128
      - Positive and Negative Patterning 128
  - Results and Discussion 129
    - Pretraining 129
    - Patterning 130

- **Study 7:** Negative and Positive Patterning with Null Trials 132
  - Method 132
    - Networks 132
    - Input Coding 132
    - Training 132
      - Pretraining 133
      - Positive Patterning 134
      - Negative Patterning 134
  - Results and Discussion 134

- **Study 8-A:** Should Networks be Losing Momentum? 138
  - Connectionist Learning and Momentum 139
  - Method 141
    - Networks 141
    - Input Coding 142
    - Training 142
      - Positive Patterning 142
      - Negative Patterning 143
  - Results and Discussion 143

- **Study 8-B:** Momentum and the Course of Learning 148
  - Method 148
  - Results and Discussion 149

- General Discussion 154

Chapter 7: GENERAL DISCUSSION 157

- Summary of Research 158

- Implications for Associative Learning Theory and Connectionism 163

REFERENCES 166

List of Tables

Table 4-1. Selected hidden unit activations for negative patterning in Study 2........ 81

Table 5-1. Means and Standard Deviations for Studies 3-A, 4 and 5-A.................. 95

Table 5-2. Hidden unit activations for the "distributed" network in Study 5-C....... 113

Table 5-3. Hidden unit activities of the four hidden units in the network in the "local" condition in Study 5-C...114

Table 6-1. Means and standard deviations for Study 6... 129

Table 6-2. Cell means and standard deviations for Study 7.....................................135

Table 6-3. Means and variances for Study 8-A..144

Table 7-1. Summary of Studies 1 - 8-B at a glance...158 - 160

List of Figures

Figure 3-1. A sigmoidal activation function. This type of function makes a single discrimination as indicated by the shadow under the function........ 30

Figure 3-2. A gaussian activation function, as used in value units (Ballard, 1986). This function makes a non-linear discrimination as shown in the shadow under the function........ 31

Figure 3-3. Rosenblatt's original conception of a perceptron........ 34

Figure 3-4. A perceptron modified as a two-layer network with no intervening units........ 35

Figure 3-5. A multilayer perceptron........ 36

Figure 4-1. Simulation Study 1: A PDP model that simultaneously learns the negative and positive patterning problems. Units A - D are used to train the positive patterning task; Units E - H are used to train the negative patterning task; Units X and Y are background stimulus cues (see text for details)........ 71

Figure 4-2. Results for Simulation 1, negative patterning task, with no pretraining of the elements. Notice that the training proceeds in a straightforward manner and that no excitatory summation is present during training........ 74

Figure 4-3. Results for Study 1, negative patterning task, with pretraining of the elements. Notice that there is excitatory summation in this condition........ 75

Figure 4-4. Network used in Study 2. A and B are elements used in negative patterning training; C and D are elements used in positive patterning training; X and Y are background stimuli (see text for details)........ 77

Figure 4-5. The course of learning for the positive patterning schedule in Study 2........ 78

Figure 4-6. The course of learning during the learning of the negative patterning schedule........ 79

Figure 5-1. Neural networks as function approximators: When functions learned in Situation "a" are very different than functions learned in Situation "b", loss of fit with data points in Situation "a" is likely.......... 89

Figure 5-2. Network used in Study 3.......... 93

Figure 5-3. Box plot identifying an outlier in the 2-Phase condition of Study 3-A.......... 97

Figure 5-4. SSE plotted against number of epochs for the 6 networks in the two inputs, two hidden units, 3-phase condition, Study 3-B.......... 99

Figure 5-5. SSE for the 4 networks in the two inputs, two hidden units, 2-phase condition, Study 3-B.......... 101

Figure 5-6. Network used in Study 4. This network is the same as that used in Study 3, except that the input patterns are coded across separate input unit pairs, depending upon the phase of training (see text for details).......... 103

Figure 5-7. SSE for the 9 networks in the local representation condition: two input units, four hidden units, Study 5.......... 110

Figure 6-1. Linear non-separability, with each discrimination represented by a line.......... 124

Figure 6-2. Training sets of Delamater et al. (1999), graphically represented. Sets in all conditions require the subjects to discriminate the elements from the compound. Given the number of inputs, this requires one discrimination which is represented by the line.......... 125

Figure 6-3. Perceptron of Study 6.......... 127

Figure 6-4. Study 6: Perceptron network, negative patterning, previously not reinforced. Box plot identifies 2 outliers.......... 130

Figure 6-5. Multilayer perceptron used in Studies 7, 8-A, and 8-B.......... 133

Figure 6-6. Study 7: No nulls, negative patterning, previously not reinforced. Box plot identifies 3 outliers.......... 136

Figure 6-7. Study 7: No nulls, negative patterning, previously reinforced. Box plot identifies 1 outlier.......... 137

Figure 6-8. Study 7: With nulls, positive patterning, previously not reinforced. Box plot identifies 1 outlier 138

Figure 6-9. Study 8-A: No nulls, positive patterning, previously not reinforced. Box plot identifies one outlier 145

Figure 6-10. Study 8-A: Without nulls, negative patterning, previously not reinforced. Box plot identifies 2 outliers 145

Figure 6-11. Study 8-A. Without nulls, negative patterning, previously not reinforced. Box plot identifies 1 outlier 146

Figure 6-12. Study 8-A: With nulls, negative patterning, previously not reinforced. Box plot identifies 1 outlier 147

Figure 6-13. Data reproduced from Delamater *et al.* (1999). Negative patterning acquisition for a group of networks learning with momentum, without null trials using previously reinforced elements 149

Figure 6-14. Data reproduced from Delamater *et al.* (1999). Negative patterning acquisition for a group of rats using previously reinforced elements 150

Figure 6-15. Negative patterning acquisition for a group of networks learning without momentum, with null trials using previously reinforced elements 150

Figure 6-16. Data reproduced from Delamater *et al.* (1999). Negative patterning acquisition for a group of networks learning with momentum, without null trials using previously *not* reinforced elements. 151

Figure 6-17. Data reproduced from Delamater *et al.* (1999). Negative patterning acquisition for a group of rats using previously *not* reinforced elements 152

Figure 6-18. Negative patterning acquisition for a group of networks learning without momentum, with null trials using previously *not* reinforced elements 153

List of Abbreviations

A	Some event, A (or substituted letter)
-A	Not A (or substituted letter)
A-, A^0	Some event, A (or substituted letter), not followed by reinforcement
A+	Some event, A (or substituted letter), followed by reinforcement
CR	Conditioned reflex
CS	Conditioned stimulus
min	Minutes
MLP	Multi-layer perceptron
n	Number of subjects
O	Outcome
P	Probability
PDP	Parallel distributed processing
R-	Response not followed by reinforcement
R+	Response followed by reinforcement
ref	Reinforcer
s.d.	Standard deviation
S-R	Stimulus-response
SSE	Sum of squared error
U_{AB}	Unique cue for compound stimulus, AB
UR	Unconditioned reflex
UCS, US	Unconditioned stimulus
X-OR	Exclusive or

Chapter 1

INTRODUCTION

Memories, world knowledge, organized thought, communication – many of the processes that allow organisms to function meaningfully – are dependent upon learning. The concept of learning is a simple one in some respects. Everyone has some lay definition of *learning*. Arriving at an operational definition has proved more difficult. Is learning a change in behaviour? Is it a change in mental state? Not surprisingly, definitions of learning have been influenced by changing assumptions in the field of psychology.

While early explorations of learning phenomena were guided by a strong adherence to the observational method and animal models, learning today is more likely to be characterized by cognitive principles. Some researchers have used the cognitive psychology assumption, that cognition is information processing, to make methodological advancements in comparative research. Traditional animal models of learning are being or have been replaced or extended by computer simulations. Some of the early computer models of associative learning have been supplanted by more sophisticated parallel distributed processing models (PDP models, neural network models, or connectionist models; Rumelhart & McClelland, 1986) that go beyond the analysis and interpretation of learning data in that they, themselves, learn.

The move toward computer models shows a willingness on the part of learning researchers to move beyond the methodology that pervades the rich history of traditional approaches to learning. However, there are theoretical implications for current conceptions of associative learning with the acceptance of these computer models as appropriate for investigations of learning phenomena. This thesis is an exploration of these theoretical implications. In the following chapters, I argue that connectionism can make contributions to learning theory that extend beyond methodology into theory, and that decisions made by researchers about the uses of these models can have a strong impact on the behaviour that these models produce and the theories that they suggest.

Associative learning has a history of good experimentation and solid progress in the understanding of learning phenomena. I propose that information can flow in both directions – with associative learning theorists gaining insight from connectionist researchers and connectionists benefitting from many of the principles of learning that have been elucidated through traditional investigations. The format of this thesis is a pattern that can be used to advance theories of learning, first by exploiting connectionist networks for insight concerning theoretical issues in the area of discrimination learning and second, by implementing known principles of learning to improve connectionist models or to improve the manner in which these models are evaluated.

In Chapter 2 a selective survey of behavioural learning theory and cognitive learning theory is presented. I explain the move from an early assumption of learning (that two things must co-occur to become associated) to the conception of associative learning that is assumed today through the work of Rescorla, Wagner and others in the late 60's and early 70's. I discuss cognitive conceptions of learning from the perspective of the mentalists in the behavioural learning tradition. The notion that cognition is information processing will be introduced, that has provided impetus for the computer metaphor. The computer metaphor will be drawn upon as a basis for the theoretical and simulation work presented in the remainder of the thesis.

In Chapter 3, I introduce connectionist models. I address the relationship between traditional associative learning theory and connectionist models, and I discuss the theory of David Shanks (1995) who adopts connectionist modeling as a research method. I suggest that, although connectionist models can be used methodologically, they are more than a methodology, as they can be used for theory development and as theories themselves. Where they are used as a method for the elucidation of traditional theories, it is appropriate to consider the implications of this decision. These implications are explored in Chapter 3.

Chapter 4 is primarily a theoretical argument that is an example of the first way that I suggest that theory can be advanced: by exploiting connectionist networks for theory construction. I describe both the limitations of early connectionist models, and the emergence of solutions to these problems and relate them to current limitations of theories of associative learning. Specifically, I demonstrate particular advantages of connectionist architectures in terms of the representation of compound stimuli (distributed representations and coarse coding) and suggest that these properties of connectionist networks be used to increase the representational power of models of discrimination learning. Two studies are included that demonstrate the computational and representational power of a relatively simple neural network model that solves a particular discrimination learning problem. An interpretation of one of these networks is provided that has implications for discrimination learning theory.

While I advocate a move toward distributed representations in Chapter 4, I relate in Chapter 5 one of the limitations of models that contain distributed representations: these networks may be particularly prone to a phenomenon called catastrophic *forgetting*, in that a network trained on Task 1 *forgets* that task after being trained on Task 2. The studies presented in Chapter 5 address this issue. If networks are prone to catastrophic forgetting, it makes them improbable models of learning. In these studies, I do not find any kind of forgetting that could be called "catastrophic" when *savings* in reacquisition of Task 1 is considered, rather than network performance on trial 1 following training on Task 2 in any of the networks considered in Chapter 5. This set of studies falls under the second category cited as important in this thesis: these studies make use of known principles of learning and memory to improve the way that connectionist models are considered and tested.

Chapter 6 contains another set of studies that suggests refinements to connectionist models based on principles of learning. This study demonstrates how learning-improbable design decisions can have an effect on neural networks and can, by consequence, affect the hypotheses that these networks generate. It is illustrated that a network that is to be evaluated on the basis of its ability to produce behaviour that fits experimental data is only one of a family of connectionist models that could be used in a given situation. A model for testing connectionist theories is presented in which a simplest model is considered first and other members of the family are considered when the network does not produce the expected behaviour.

Further results and conclusions of the above studies are discussed in Chapter 7. Through the thesis, evidence is provided for the proposition that connectionist models can provide interesting and powerful extensions to theories of associative learning, when they are used with careful consideration. Arguments are made throughout the thesis, for thoroughly considering the implications of using connectionist networks in learning research, and for considering the implications of various design decisions that are made when networks are used in learning research and theory.

CHAPTER 2

THE LEARNING OF ASSOCIATIONS

In this chapter, I present a brief review of principles of learning, first from a behavioural, then from a cognitive perspective. To prepare for a discussion of connectionist associations in Chapter 3, I will outline the work of some who have made contributions to our current conception of associative learning in this chapter. Theorists and theories have been included because of their relevance to the subsequent chapters in this thesis; this history is not meant to be exhaustive.

In the first part of this chapter, two underlying assumptions of associative learning are contrasted. The first, *contiguity*, is the principle of temporal pairing (Shanks, 1995). Contiguity-based theories assume that associations are formed when two events co-occur or occur closely in time. Contiguity can easily be demonstrated in the laboratory by implementing an operant conditioning paradigm, for example, in which a pigeon is reinforced with food for pecking a key after a specified delay. It has been reliably shown that, as the delay between the pecking and the reinforcement increases, responses are fewer and less likely. Conversely, when reinforcement is delivered immediately after pecking, learning is rapid. Contiguity is, indeed, important for this type of learning.

Contiguity theories were pre-eminent in behavioural accounts of associative learning until the late 1960's when theorists demonstrated that contiguity could not account for all known phenomena. Contiguity is necessary but not sufficient for learning to occur. *Contingency* was introduced as an extension of and an alternative to contiguity.

Contingency also assumes temporal pairing but assumes directional dependence as well: the occurrence of one event is *contingent* upon the occurrence of another event. In classical conditioning terms, contingency requires that the probability of the unconditioned stimulus, given the conditioned stimulus, is greater than the probability of the unconditioned stimulus, given the absence of the conditioned stimulus [P(US/CS) > P(US/no CS)]. In operant conditioning terms, contingency requires that the probability of the delivery of the reinforcer, given the operant response is greater than the probability of no delivery of the reinforcer, given the operant response [P(ref/CR) > P(ref/no CR)]. Underlying contingency theory is the concept of correlation: the learner computes (by some method) the correlation between two events (Shanks, 1995).

This chapter moves from early and modern contiguity theories to contingency theory, in particular, the Rescorla-Wagner model, since it has a well known and strong relationship to learning in connectionist networks. The final section describes the influence of the cognitive revolution on psychologists' perception of learning and introduces the concept of representations that will be revisited in Chapters 3, 4, and 5.

Behavioural Theories of Associative Learning: Contiguity

Learning and memory are intricately linked and fully dependent. The relationship between learning and memory is mediated through representations, that will be considered later in this chapter. As such, while many would begin a history of the formal analysis of learning with Pavlov and Thorndike, I begin with Ebbinghaus, the memory researcher. Ebbinghaus tested memory for lists of learned non-words and was, through testing memory, empirically testing the process of learning long before Thorndike.

Ebbinghaus

Ebbinghaus started his learning and memory work in 1879 and began to publish results of his studies in 1885 (Leahey, 1997). As far as our discussion is concerned, Ebbinghaus' most significant contributions were his rejection of established laws of learning that had never been subjected to empirical investigation, and his proposal of a new theory of associations. Prior to the 1880's, associations had been regarded as though they had always existed: Ebbinghaus began to study associations as they were being formed. He proposed that associations between objects could be influenced, tampered with, and analyzed (Ebbinghaus, 1885).

Using himself as a subject, Ebbinghaus studied retention and savings of groups of non-word syllables that were neither semantically nor aesthetically related. He plotted the number of recall errors as a function of number of exposures to the group of non-word syllables, creating the first *learning curve* – a staple of experimental psychology since. Ebbinghaus formulated a principle of learning, based on his study of associations. This was the *law of frequency*: the more frequently an experience occurs, the more easily it is recalled. The law of frequency re-emerges in later theories of learning, in a somewhat more complex form. Although Ebbinghaus predated contiguity theory, the law of frequency becomes an important part of the temporal pairing hypothesis: that the more *frequently* two events co-occur or occur close in time, the stronger the association between them (Shanks, 1995). This is the birth of the science of Associationism – the notion that "all knowledge is based upon the associations between ideas" (p. 5).

In spite of his empirical, atheoretical approach (Leahey, 1997), Ebbinghaus retained some notions about learning that were considered antiquated and non-scientific. He believed learning was the acquisition of logical relations between ideas – a very mentalistic concept that fell out of favour with learning theorists until much later in the 20th century, when it made a comeback as a basis for associative learning and for physiological theories such as Hebb's (1949).

Ebbinghaus engaged in research inspired by an interest in human higher mental processes. The Russian physiologist, Pavlov was, at approximately the same time, studying the cerebral cortex in animals from a physiological perspective. These perspectives appear to have little in common. However, the vastly different theories and methodological approaches of these two researchers have, together, provided what many would consider the original foundation of learning theory and research.

Pavlov

Ivan Pavlov began investigating learned or conditioned reflexes in dogs about 15 years after Ebbinghaus published his short monograph, *Memory* in 1885. Pavlov was critical of existing methods of psychology claiming that "if we attempt an approach from this science of psychology to the problem confronting us we shall be building our superstructure on a science which has no claim to exactness as compared even with physiology" (Pavlov, 1927, p. 3).

Pavlov's mechanistic view was influenced by Descartes, who asserted that every activity of the organism was a reflex (i.e., an obligatory reaction to a stimulus from the environment). Pavlov found that some activities of organisms, however, did not work like *simple* reflexes. He could demonstrate complex reflexes by establishing a setting in which a contingency existed between a stimulus (like a tone) and an outcome (like meat powder delivered to the mouth of a hungry dog). A simple reflex to meat powder in the mouth is salivation and, certainly, Pavlov found this simple reflex in his subjects. He called the meat powder the unconditioned stimulus (US) and the salivation to the meat powder the unconditioned reflex (UR). When a tone was reliably sounded prior to the administration of the meat powder, the dogs began to salivate at the sound of the tone that previously had no power to elicit this response. Pavlov called the tone the conditioned stimulus (CS) and the salivation to the tone the conditioned reflex (CR; Pavlov, 1927).

To explain this phenomenon, Pavlov formulated the principle of *stimulus substitution*: a reinforcer elicits a behaviour that comes to be associated with any stimuli that occur closely in time with the reinforcer. In the example above, the CS (the tone) comes to represent the US (the meat powder) and, hence, the response is elicited by the CS because of a learned equivalence between the US and the CS. In later learning terminology, *stimulus-stimulus* associations are formed. As a physiologist, Pavlov postulated that US-UR associations are innate and that CS-US associations came about as the result of the strengthening of neural connections.

Pavlov went on from the simple scenario of the tone and the meat powder to discover many interesting phenomena associated with classical conditioning. In his 1927 book *Conditioned Reflexes*, Pavlov documented phenomena that continue to generate research today such as secondary conditioned reflexes, habituation, extinction, overshadowing, spontaneous recovery, conditioned inhibition, inhibition of delay, and responses to compound stimuli.

Although some of the phenomena outlined in *Conditioned Reflexes* require more complex theory (e.g., some compound stimulus tasks, overshadowing), the underlying basis of Pavlov's work was temporal contiguity: when the CS and the US occur closely in time, the US comes to represent the CS that elicits a UR: conditioning occurs.

While Pavlov was documenting the results of his many studies in Eastern Europe, similar work was being undertaken in America in the Harvard University lab of Edward Thorndike. While Pavlov's empirical work predates that of Thorndike, Thorndike began *publishing* his work before Pavlov published the Russian version of *Conditioned Reflexes* (Leahey, 1997). Both researchers studied conditioning but their work overlapped little. Pavlov was interested in the conditioned reflex; Thorndike became a specialist in a type of learning presumed, at the time, to have a different underlying mechanism – operant learning.

Thorndike

In initial experiments, Thorndike placed a cat in a cage-like "puzzle box" that had a pull string that acted as a release mechanism for the cage door. Typically, a cat would struggle to escape the cage until it accidentally pulled the escape cord. The next time the cat was placed in the puzzle box, it was more likely to pull the string. It had *learned* that pulling the string would result in escape from the box (Leahey, 1997).

Thorndike initially explained this learning with the *law of effect* and the *law of frequency*. The law of effect states that the connection between a stimulus (the puzzle box) and a response (pulling the string) is strengthened if the response is followed by a satisfying outcome (getting out of the box) or weakened if the response is followed by an aversive state of affairs. The law of frequency, according to Thorndike, is that the more often a situation connects with a particular response "the stronger becomes the tendency for it to do so in the future" (1932, p. 6). His early theory included other laws such as the "law of readiness" and the "law of exercise". By his 1932 book, *The Fundamentals of Learning*, Thorndike had reduced the importance of many of his secondary laws and had substantially revised the law of effect (Bolles, 1975). Only the first half of the law of effect could be substantiated (the connection between a stimulus and a response is strengthened if the response is followed by a satisfying outcome); the punishment side of the law of effect could not be corroborated.

The law of effect is important to the present discussion because it served as a break from existing concepts of the nature of associations in learning. Formerly, mere contiguity or temporal pairing was presumed to be sufficient for conditioning. The law of effect made it clear that the strength of the stimulus-response (S-R) connection is determined by the consequences of the response. Although Thorndike's concept of the reinforcer is added to contiguity to produce the law of effect, Thorndike's theory does not overwrite contiguity theory. Rather, for Thorndike, "the critical feature of the response-reinforcer relation was temporal contiguity" (Schwartz & Robbins, 1995, p. 208).

Thorndike's approach was mechanistic. Although the organism was presumed to hold an idea that was connected (weakly or strongly) to another idea, it was not the "insight" or "logical relations" of earlier theories. Thorndike argued that if learning were the outcome of insight, then the learning curve that Ebbinghaus had described would be more like a learning cliff – something that was neither empirically nor logically plausible.

Thorndike's theory came to be known as *Connectionism*, named after the bonds or associations between sensory input and behaviour (Hilgard, 1956).

For some, however, the concept of the reinforcer was still too mentalistic. The relevance of the reinforcer, as introduced by Thorndike, became a matter of debate in psychology. The American psychologist who most successfully challenged the importance of the reinforcer was John B. Watson who had a limited time in the psychology spotlight, but a long-lasting effect on American psychology (Bolles, 1975).

Watson

Watson's major departure from many who came before was concerning the relevance of the reinforcer. Watson claimed that for the reinforcer to have an effect on the S-R connection, the organism would have to anticipate the reinforcer – a mentalistic concept. To avoid this non-observable explanation, Watson stressed the law of frequency: that the more often a stimulus and a response occur together, the stronger the connection between the two will become.

But reinforcers appear to work! After all, animals come to respond at higher rates when responses are reinforced. Watson explained this phenomenon by describing a hypothetical experiment in which there are two possible responses. One response (R-) never results in reinforcement while the other response (R+) always does. A single trial finishes when the organism makes the R+ response. Watson points out that although initially, the organism should respond with R- and R+ equally, each and every trial must contain an R+ in order to terminate but need not contain an R-, therefore R+ is more frequently paired with the situation/stimulus (Bolles, 1975).

Like Thorndike, Watson asserted that when stimulus and response occur at the same time (or close in time), the connection between them is strengthened. Watson also promoted the idea that, neurologically speaking, connections are not created through learning – rather the neurological connections are present prior to the pairing of a stimulus and response. The pairing either lowers the threshold of the old connection (making it more likely to "fire") or awakens a latent connection (Bolles, 1975). Watson's theory is another example of a contiguity-based theory; at its most basic level, this theory is about the temporal pairing of stimulus and response.

Other aspects of Watson's theory have not stood up well in the face of empirical evidence. Watson is interesting to this discussion, however, for his influence on American psychology that endures to the present. Although he did not invent *behaviourism* as Thorndike's theory must also be considered *behaviouralist*, Watson is considered the first *Behaviorist*. His absolute rejection of the mentalistic perspective has persisted in learning research and in many theories of learning, until relatively recently.

The influence of behaviourism is clear in many subsequent theories of learning. There were, however, many versions of behaviourism besides Watson's well-known theory (Hilgard, 1956). Edwin Guthrie, for example, devised his own behaviourist position inspired primarily by the views of the philosopher Singer (Bolles, 1975). Guthrie's perspective differed in some key ways from the behaviourism of Watson but many concepts present in both Watson's and Thorndike's theories surfaced also in Guthrie's theory of learning.

Guthrie

Guthrie's theory of stimulus substitution contains echos of theories already discussed but is sufficiently distinctive to warrant its own discussion. Like Watson, Guthrie omitted the reinforcement mechanism from his theory, believing it to be mentalistic. He also stressed the law of frequency, that he called "positive adaptation." Like Pavlov did for the conditioned reflex, Guthrie claimed that operant learning was a function of stimulus substitution (that he called "conditioning") in which the CS comes to substitute for a US in eliciting a response (Guthrie, 1935). He stressed temporal pairing as being important in building the CS-US association.

One of the unique aspects of Guthrie's theory was his argument that learning is instantaneous. Upon presentation of a particular pattern of stimuli that was followed by a particular response, an organism could be instantly conditioned to that stimulus pattern. He explained the gradual shape of the learning curve by pointing out that stimuli are part of an ever changing situation. From one trial to the next, subtle aspects of the stimulus situation, the subject, and the response are likely to change. The organism does not respond to the stimulus that the researcher may believe he or she is presenting; it responds to a stimulus pattern. Guthrie explained this:

> The psychologist must resign himself to the fact that no psychological event is ever really repeated. The second repetition of a stimulus is only roughly and for practical purposes equivalent to the first; his laboratory subject is only substantially or approximately the same person who sat in the chair the day before . . . no two responses are alike. Two trips through a maze, two conditioned

> salivary reflexes may be substantially the same, but they are always the same with a difference. (1935, pp. 10-11)

In keeping with Guthrie's assertion that the consequences of a particular response (the reinforcer) were inconsequential to learning, he explained why reinforcement appears to work with the following scenario: An organism is placed in situation *A*. Every time the organism makes a slight response, the situation changes slightly, but is still recognizable as situation *A*. When situation *A* changes slightly, the organism "writes over" the stimulus pattern-response association for situation *A*. This continues until the organism makes a response that changes situation *A* to situation *B*, for example, when the "appropriate" response takes place and the situation changes to one in which something is presented to the organism (a reinforcer). When situation *B* is entered, the organism no longer writes over the situation *A* association. Therefore, the stimulus pattern-response association, that has been reinforced, is maintained.

Guthrie's theory is, again, an example of a contiguity-based theory. Although his theory appears quite unique, the foundation on which his theory is built is still contiguity: temporal pairing builds associations between stimuli. Like Watson's theory, this is also a mechanistic theory. The behaviourism of Thorndike, Watson, and Guthrie was indeed more mechanistic and was considered more scientific than the philosophies of mind that predated them. In isolating the observable, these theorists aspired to create a science of behaviour.

By the 1930's, this science of behaviour was ready for an in depth, quantitative, and formal analysis of behaviour. Clark Hull entered the field in 1920. In the next 20 years, Hull's commitment to a formal model of learning yielded a series of influential concepts, and culminated in the publication of *Principles of Behaviour* in 1943 (Leahey, 1997).

Hull

Clark Hull was a synthesizer of theories that had come before. His general theory was similar to Thorndike's in that the S-R connection required reinforcement in order to be strengthened. He believed that the CS was substituted for the US, as did Pavlov. His theory was similar to that of Edward Tolman who claimed that animals behave purposefully toward goals that are adaptive. Hull did not shy away from terms such as "expectancy" but claimed to be neither mentalistic nor mechanistic – instead, he described himself as behaviouristic, after Watson. He went beyond former theories, however, with his quest for a quantitative model of behaviour (Bolles, 1975).

Hull also stressed that drive reduction is crucial for reinforcement. An organism must be motivated by a biological drive in order to behave. He distinguished between primary reinforcers that are direct biological drive reducers and secondary reinforcers that have been conditioned as reinforcers (e.g., social approval for humans). A drive, such as hunger, creates general behaviour – not just behaviour related to the drive state. Eventually, an organism is rewarded for a particular behaviour – pecking at food if the drive state is hunger, for example. That behaviour becomes more likely again when the organism enters into *any* drive state (Hull, 1943).

Hull reformulated Thorndike's law of effect according to his drive-reduction theory:

> Whenever an effect or activity occurs in temporal contiguity with the afferent impulse, or the perseverative trace of such an impulse, resulting from the impact of a stimulus energy upon a receptor, and this conjunction is closely associated in time with the diminution in the receptor discharge characteristic of a need, there will be an increment to the tendency for that stimulus on subsequent occasions to evoke that reaction. (1943, p. 80)

Hull's theory clearly remains within the framework of contiguity-based theories.

Although Hull's drive-reduction theory has not been supported by research (for example drive states motivate specific rather than general behaviour) many of his ideas have stimulated further research and his mathematical formulations have encouraged formal theory within the discipline of learning (Bolles, 1975).

Summary

Early theoretical differences in classical conditioning were primarily of the micro variety: Are connections made between the unconditioned stimulus and the conditioned stimulus, or between the conditioned stimulus and the unconditioned response? It has since been generally agreed that an association must be formed between the stimulus and the response to that stimulus, however, Mackintosh (1974) argued that "there is no *a priori* reason why animals should not associate a CS both with a UCS and with their reaction to that UCS" (p. 89). In fact animals may be making many associations that affect the likelihood of behaviours. Operant conditioning is assumed to be mitigated by the formation of an association between a response that has been reinforced and the situation in which the reinforcement occurs. According to Mackintosh, "The most general associationist view would hold that animals may associate any set of events that happen to be correlated in time" (1974, p. 140).

After the work of theorists like Hull, emphasis moved from the location of the associations to the nature of those associations. Theoretical differences had centred around *what* things became associated when a stimulus and a response occur closely in time. The question that moved to the forefront in the late 1960's was whether two things occurring closely in time was sufficient to produce conditioning (Rescorla & Wagner, 1972; Rescorla, 1968). This question led to two outcomes: a reformulation of contiguity theory and an alternative to contiguity: contingency.

The reformulation of contiguity theory suggests that, rather than absolute temporal contiguity being the basis for associative learning, *relative* estimates of temporal contiguity may be important for learning. Gibbon and Balsam (1981) suggest that an organism evaluates the CS-US interval (the time between the onset of a CS and the delivery of a US) relative to the time between USs. If the CS-US interval is shorter than the US-US interval in the absence of the CS, the CS becomes associated with the onset of the US (see also Jenkins, Barnes & Barrera, 1981).

Relative contiguity theory differs from absolute contiguity theory in that, while absolute contingency requires only the co-occurrence of two events or that two events occur close to each other in time, relative contiguity requires that one event is more likely, given another event. Relative contiguity then, is a close relative of the notion of contingency in which *directional dependence* is presumed: one event is contingent upon another.

Behavioural Theories of Associative Learning: Contingency

Although Associationism explained basic learning phenomena well, it became apparent during the 1960's that contiguity theory could not account for more complex phenomena such as compound stimulus tasks and blocking (Rescorla, 1969; Wagner, 1968, 1969a; Wagner, Logan, Haberlandt, & Price, 1968). In 1972, Rescorla and Wagner published a paper in which they showed that contingency, not contiguity, is the condition necessary for learning to occur. Statistically speaking, contingency is "simply the calculation of the degree to which a pair of events covary" (Shanks, 1995, p. 21): contingency is an estimate of the correlation between two events. For a computation of contingency, it is important not just that A predicts B, but also that not A predicts not B (Rescorla, 1968).

Humans and other organisms certainly seem to be responsive to correlation (see Rescorla, 1968) but the computation involved in estimating degree of relatedness remains a matter of speculation. Some have proposed a χ^2 calculation in which

$$\chi^2 = \frac{N(ad - bc)^2}{[(a+b)(c+d)(a+c)(b+d)]}$$

The degree of association between two stimuli (χ^2) is predicted based on a comparison between a, the frequency of the co-occurrence of the two stimuli and d, the frequency of the non-occurrence of both stimuli and b and c, the frequency of one of the stimuli without the other. χ^2 is a covariation calculation for binary variables and is a two-way dependency (is stimulus 1 dependent on stimulus 2 or is stimulus 2 dependent on stimulus 1?). While χ^2 is an actual calculation of correlation, it ignores the contingency requirement of directional dependence.

Another proposed rule, ΔP, on the other hand, *is* directionally dependent:

$$\Delta P = P(US|CS) - P(US|no\ CS)$$

in which ΔP (which is actually a correlation calculation for continuous variables) represents change in associative strength. Both χ^2 and ΔP have had some empirical validation but, according to Allan and Jenkins (1983), no proposed rule accurately predicts human data. Allan and Jenkins found that the ΔD rule

$$\Delta D = (a+d) - (b+c)$$

most closely conforms, but still remains an inadequate ($r = .73$) explanation of the data.

The Rescorla-Wagner Model

In an attempt to model associative learning within the framework of contingency theory, Rescorla and Wagner developed a basic equation that contains some elements that had not been represented in previous models. In the Rescorla-Wagner model, learning is due to changes in associative strength between a conditioned stimulus and an unconditioned stimulus. Rescorla and Wagner assume that context is important in conditioning and they incorporate background into their model as part of a compound stimulus. For a conditioned stimulus, *A*, the following equation represents change in associative strength:

$$\Delta V_a = \alpha \beta (\lambda - V_{ax})$$

in which ΔV_a is change in associative strength, α is a learning rate parameter that is dependent upon the salience of a component stimulus, β is a learning rate parameter dependent upon the nature of the unconditioned stimulus, λ is the asymptotic value of possible associative strength, and V_{ax} is the associative strength among all active stimuli (Rescorla & Wagner, 1972). Statistically speaking, the learning organism calculates an association or covariation between pairs of events (Shanks, 1995). As the difference between the current associative strength and the maximum associative strength that can be attributed to a connection decreases, less conditioning occurs. In other words, as conditioning proceeds through trials, less is learned on each trial. This explains the deceleration of learning as demonstrated by the acquisition curve, and accounts for overshadowing (when *A* and *B* are presented as a compound stimulus to predict a US. Subsequently, when *A* is presented without *B*, there is a deficit in responding) and blocking (*A* is paired with the US. Subsequently, when *A* is presented with *B*, subjects fail to condition to stimulus *B*).

The Rescorla-Wagner model has become the dominant model of associative learning: the model against which other models of learning are evaluated. There seems to be no flagging of interest in this model. In fact, Pearce and Bouton (2001) report that between the years 1981 and 1985, the Social Science Citation Index records more than 330 citations while between the years 1995 and 1999 more than 480 citations are recorded. Given that this model is now 25 years old, it has aged remarkably gracefully and has withstood many empirical criticisms (see Miller, Barnet, & Graham, 1995 for a review of the successes and failures of the model).

Contiguity and Contingency Today

The contiguity/contingency debate resurfaces occasionally when one or the other is unable to account for a particular phenomenon. Miller & Matzel (1988) note that for an organism to calculate contingency, the organism must *know* how often the CS has occurred with the US: that is, the learning of the CS-US association must already have taken place before contingency has been calculated. Hallam, Grahame, Harris, & Miller (1992) have suggested that temporal contiguity may be responsible for the acquisition of behaviour while contingency may direct behavioural expression after learning.

Limitations of the computational power of both contiguity and contingency theories will be discussed in the following chapter.

Associative Learning: The Cognitive Perspective

While the purer behavioural theories became less and less mentalistic through the decades leading up to the 1960's, mentalism was gaining momentum among a group who became interested in learning in terms of how it could inform about what was processed and stored in the mind when something was learned: the cognitivists. The cognitive movement is generally considered a revolution (Lachman, Lachman, & Butterfield, 1979) of Kuhnian proportions (Kuhn, 1962/1970) in psychology. Cognitive psychology's most obvious ancestors are in the fields of mathematics and artificial intelligence and go much further back than "the dawn of cognitivism" in psychology in the 1970's. Although viewed as an alternative to behavioural psychology however, many of the roots of the cognitive movement *in psychology* can be traced back to those theorists of the behavioural era who considered associations between *mental* events to be important for learning.

Cognition in Behavioural Approaches

An example of an early cognitivist within the behavioural tradition is Edward Tolman. While Tolman was an Associationist in that he believed that associations are learned by temporal contiguity, his theory fell outside the mainstream of theory in his time (Bolles, 1975). He was opposed to the notion that reflexes were solely responsible for behaviour and that all behaviour was determined. Instead, he emphasized purposive, goal-directed, intelligent aspects of behaviour.

He suggested that organisms learn to predict events based on some expectancy of that event (Tolman, 1932). Tolman theorized that a pattern of many expectancies creates a "cognitive map" in which information such as temporal information relevant to many associations may be stored. These maps allow organisms to evaluate the expected outcomes associated with a number of alternative behaviours. The expectancies stored in a cognitive map are modified and governed by a set of syntactic rules that, while different in complexity, are not so different in substance from the widely accepted Stimulus-Response rule thought to govern all of behaviour by the mechanistic theorists (Bolles, 1975).

The importance of Tolman's work to the present discussion is primarily contained in the language that he used to describe his theory. He introduced some cognitive constructs into behavioural theory long before cognition had become a driving force in psychology. Some aspects of Tolman's theory, including his notion of cognitive maps and syntactic rules are recapitulated in other, later, cognitive theories of behaviour.

Interestingly, although theorist Clark Hull rejected mentalistic constructs that form the basis of cognitive theories and debated frequently with Tolman concerning mechanistic and purposive behaviour (Leahey, 1997), some of the work of Hull foreshadowed the move in cognitive circles toward the computer metaphor. Hull, who was committed to principles of mathematics and physics was also convinced that machines could think. He believed that mechanistic psychology would be fulfilled in machines that learned (Leahey, 1997). He built prototypes of intelligent, artificial systems and claimed that "learning and thought are here conceived as by no means necessarily a function of living protoplasm than is serial locomotion" (Hull & Baernstein, 1929; in Leahey, 1997). While his mechanistic philosophy remained at the forefront of Hull's psychology, intelligent machines became less important in his work as his mathematical theory became more widely accepted.

Modern Cognitive Psychology and Learning

Understanding the processes of learning was of primary importance to the behavioural theorists who could be described as mentalistic or cognitive. For the cognitivists of the 1970's, however, learning became less important as the focus of theories and research turned to *representations*.

Representations are the mental codes or traces that allow organisms to remember the results of experience. The relationship between learning and representation seems clear: that the process of learning about stimuli builds representations of those stimuli. Representations, however, have not always been easy to define. Representational systems require the assumption of two worlds: the real world that is being represented and the representational world (Roiblat, 1982). A particular representation requires a mental mapping between the real world (some sensory input) and the representational world.

Cognitive psychology also became increasingly enamored with a computer metaphor throughout the 1970's. The primary assumption of the computer metaphor is that there is a similarity between processing that is done in a computer and processing that is done in the mind. Computers and organisms receive input, process information, produce output. Central to this perspective is the notion that cognition is information processing or, at least, that cognition is *like* information processing. Within behavioural circles, the stimulus-response assumption of earlier models has, for the most part, been replaced with a stimulus-processing-response assumption, allowing for some computer-like processes in models of associative learning.

Summary

In this chapter, I have briefly reviewed some theories of associative learning. In the next chapter, I will present another approach to learning, Parallel Distributed Processing (PDP). PDP technique and theory owe much to the fundamental principles of learning mentioned above (Kehoe, 1990) and to the prime tenet of the information processing branch of the cognitive movement: that cognition is information processing.

PDP models provide the opportunity to pursue questions of learning and questions of representation simultaneously while simulating the learning of associations in information processing systems. Associations, contingency, and representations will be considered in Chapter 3 in the context of connectionist models of learning.

Chapter 3

CONNECTIONISM: LEVELS OF ANALYSIS

As has been shown in Chapter 2, investigations of learning phenomena have traditionally taken place in the labs of researchers under the umbrella of psychology. Theories of learning were drafted by scientists who were trained as psychologists and who initially reduced the vast problem space associated with learning by looking at animal models. More recently, cognitive explanations of learning have become mainstream in psychology, and researchers in other areas have become interested in learning phenomena. While theories of associative learning were being developed, alternate lines of inquiry were concurrently being pursued outside of the discipline of psychology: the information-processing theories of cognition and neuroscience explanations of brain physiology have been explored together in *connectionist* circles.

While connectionism has not generally been considered in reviews of associative theory, PDP models have properties that can inform learning theory. Connectionist models are, in the most general sense, associative in nature (Bechtel, 1985). In this chapter, I describe the basic building blocks of connectionist architectures and their relationship to theories of associative learning. I will also outline the argument of David Shanks, a learning theorist and researcher who uses the tri-level hypothesis of David Marr (1982) – a framework for research adopted in cognitive science – to explicate the state of associative learning theory today. I then suggest an alternate way of viewing connectionist models of learning phenomena, more in line with most of the proponents of PDP modeling: as more than simply an implementation of existing associative learning theory. This perspective takes advantage of the particular properties of connectionist models that makes them powerful models of learning and cognition.

Basic Principles of PDP Models

Connectionist models are information processing systems. The basic principles of connectionist modeling are relatively simple, although there are numerous variations. Unlike standard computer programs that respond to input in exactly the way the programmer has programmed them to, connectionist networks can *learn* how to classify patterns or instances. Connectionist networks are made up of a number of elements that allow them to process input in a brainlike fashion (Dawson, 1998). These elements include the basic building blocks of connectionism: the processing units.

Processing Units

Neuronal inspiration. Connectionist systems are built of processing units, that are *analogous* to neurons.[1] A processing unit simulates a neuron in that it processes an electronic signal from a number of sources, and attenuates the signal through synapses to adjacent units. In connectionist networks, the processing units are arranged in layers and are connected to one another via weighted connections. These weighted connections are generally described as a simulation of a synapse.

Activation functions. Processing units compute net input, then adopt an internal activity level according to an activation function. The most basic of these is a *binary function* or a step function in which the strength of the output signal is 0 when the sum of

---

[1] The neural analogy is explored later in this chapter.

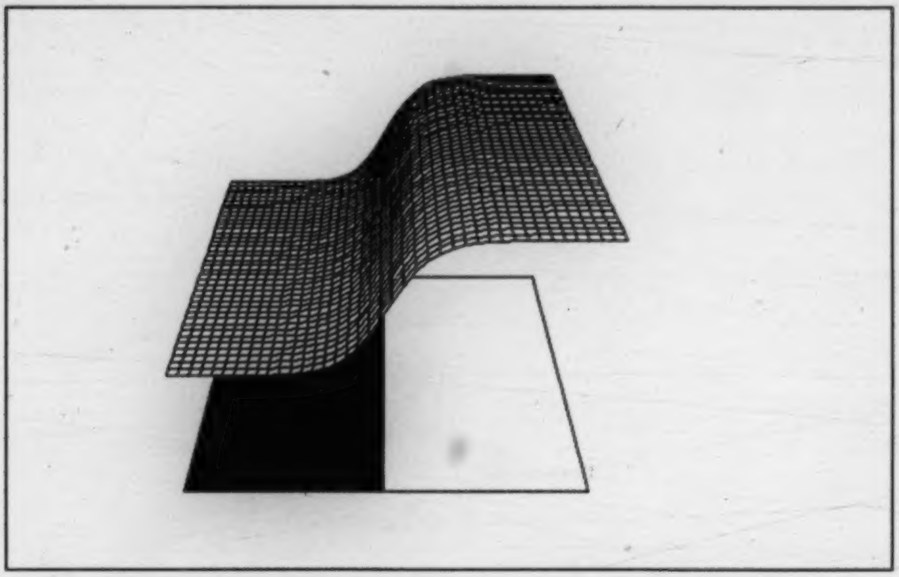

Figure 3-1. A sigmoidal activation function. This type of function makes a single discrimination as indicated by the shadow under the function.

the inputs is less than a threshold value and 1 when the sum is greater than that value. Processing units with binary functions are limited to linear classifications.

Another common activation function is a *sigmoidal function* in which the signal is "squashed" to approximate a binary function while allowing intermediate signals to be adopted, rather than limiting activity to 0 or 1. The sigmoidal function transforms negative inputs to a positive value less than .5, a net input of 0 into an internal activity of .5, and a positive input into a value between .5 and 1.00. Like the binary function, the sigmoidal activation function is a monotonic classifier (see Figure 3-1).

In some networks, a non-monotonic activation function, such as a *Gaussian function* is used. The Gaussian transforms high or low net inputs to 0 and mid-range net inputs to 1. This enables the processing unit to make two linear discriminations (see

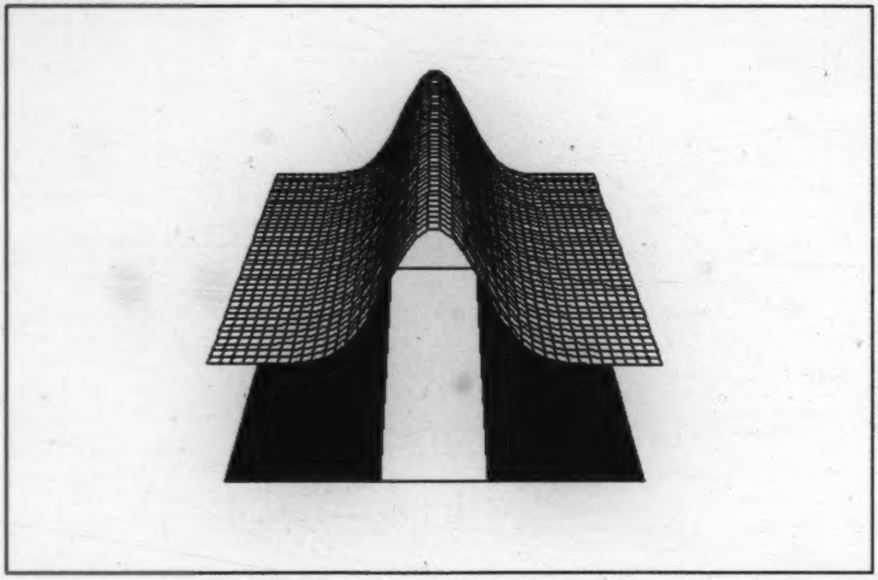

Figure 3-2. A gaussian activation function, as used in value units (Ballard, 1986). This function makes a non-linear discrimination as shown in the shadow under the function.

Figure 3-2). Processing units with non-monotonic activation functions are sometimes referred to as *value units* (Ballard, 1986). Although non-monotonic activation functions have been recognized as a theoretically powerful possibility, they have not traditionally been used. Recently, a method for training networks of value units has been developed (Dawson & Schopflocher, 1992).

The processing units in a connectionist system are combined into an interconnected network. They are organized into discrete layers in which input is processed. The combination of processing units into a neural network is a matter of network topology.

Network Topology

The first layer of processing units in the network are *input units* that are most usually either turned *off* or *on* to represent a pattern or a problem. Input units send information on in the network via connections between the first layer and the next layer.

That next layer is made up of *hidden units* that are processing units that receive a signal from each of the input units. They calculate an internal level of activity, based on the net input, then relay a signal to the next layer of processing units. Hidden units *pre-process* (Bechtel & Abrahamsen, 1991) information by detecting features in the input patterns that aid the network in learning an appropriate *mapping* of associations between the input space and the output space.

The final layer in a network is made up of *output units*. Output units are processing units that receive a signal from the hidden units. They compute the total signal being sent and then adopt a level of internal activity. In some networks, output units receive a signal only from the hidden units, while in other networks they are also connected directly to the input units.

Regardless of the network topology or the activation function of the processing units, all networks *learn*. Learning is generally accomplished through the same mechanism that governs learning in traditional models of learning: through experience. In the case of PDP models, a network changes its representation of knowledge as new information is acquired.

Learning in PDP Networks

Knowledge is represented in a network in the connections between processing units. The processing units in each layer of a connectionist system are generally connected to all processing units in the next layer. The connections between the units are modifiable and are weighted. A signal moving through a network passes through these connections; a weight attenuates the signal. It is the pattern of the connection weights that determines what the network knows and how a particular pattern is classified.

When a network is being trained on a particular problem, a set of input patterns that represent the problem is presented to the network and connection weights are modified according to rules. Generally, a network is informed of the difference between its output and the expected output – in cognitive science terms, the network is *supervised.* The network adjusts its connection weights to account for the error and the process is repeated as another set of inputs is presented to the network until the network performs the task to some specified degree of accuracy. Most often, feedback is given to the network concerning how much its output differs from the expected output and weights are changed enabling the network to more correctly classify the patterns. A set of patterns is presented repeatedly until some level of accuracy has been achieved by the network. Each presentation of a complete set of training patterns is called an *epoch* or a *sweep*. The modification of the weights in a network is the key to how the network learns. This modification is done according to a learning rule.

As all the building blocks of PDP networks can be varied, there are a diverse array of networks that can be classified as PDP. To illustrate the basic structure of PDP networks, and to lay the foundation for a discussion on learning within networks, I will briefly describe two types of networks, the perceptron and the multi-layer perceptron.

Sample PDP Networks

Perceptrons

In 1958, Rosenblatt described a network of units called a perceptron that, theoretically, is a powerful model. In Rosenblatt's network there are sensory input units, intervening association units, and output units, linked together via weighted connections (see Figure 3-3). Not all intervening units are linked to all input units; the connections are randomly assigned. Likewise, not all output units are linked to all intervening units. The processing units in the perceptron introduced by Rosenblatt have binary activation functions.

Practically, however, there was no method for training this type of perceptron

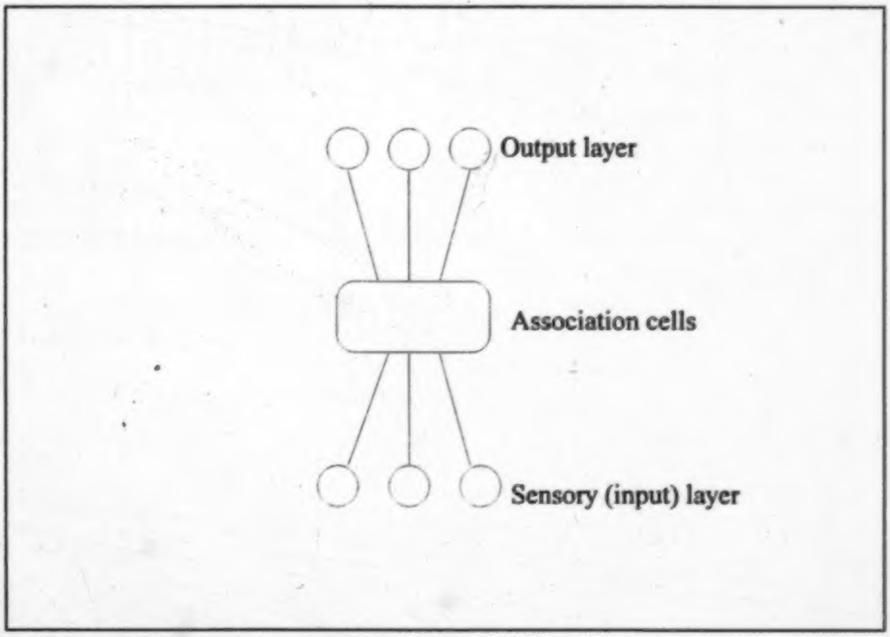

Figure 3-3. Rosenblatt's original conception of a perceptron.

when it was introduced. A simplified, trainable version of this theoretical perceptron was devised. This trainable network is different from Rosenblatt's original conception of a perceptron in that the input units are directly linked to the output unit and no intervening layer is allowed. This type of network has inherited the name "perceptron" from Rosenblatt's theoretical network, in spite of this obvious difference. Hereafter, the term "perceptron" is used to define the two-layer network with no intervening units.

The basic form of a perceptron is a simple network that only contains input units and a single output unit (see Figure 3-4). Generally, the processing units in a perceptron have binary activation functions. Simple perceptrons are defined as networks with input units being connected to a single output unit, with no intervening processing units, and no

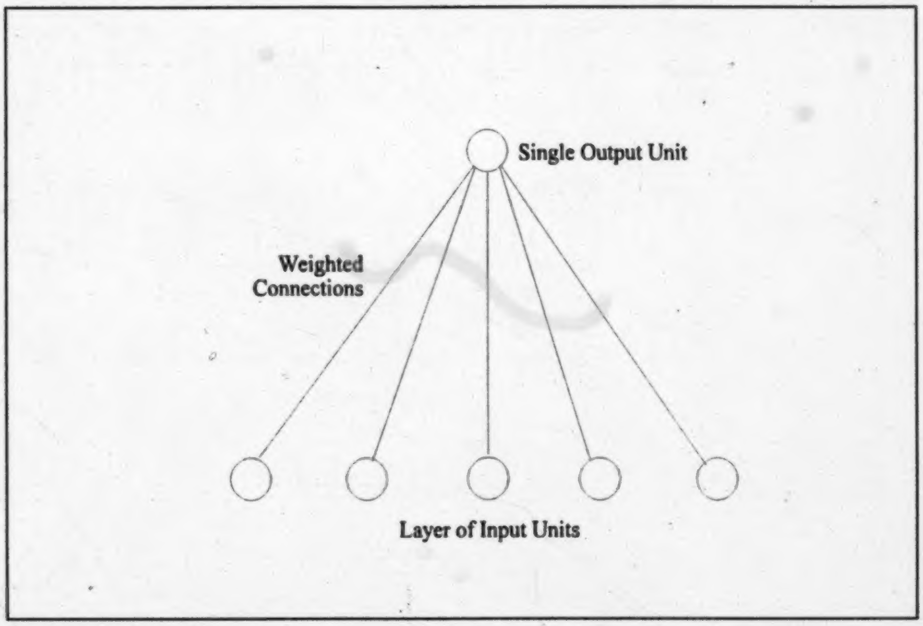

Figure 3-4. A perceptron modified as a two-layer network with no intervening units.

other connections present. Perceptrons can learn to classify patterns by adjusting their connection weights according to a learning rule.

Multi-layer perceptrons

In the 1980's, methods were developed for training networks more complicated than simple perceptrons. A more complex version of the perceptron, multi-layer perceptrons (Rumelhart, Hinton & Williams, 1986) have layers of interconnected processing units, similar to the proposed structure of Rosenblatt's original perceptron. A layer of hidden units (that each receive weighted input from each input unit) compute an activation and send a signal to the next layer of processors, is added to a perceptron (as in Figure 3-5), making a multi-layer network.

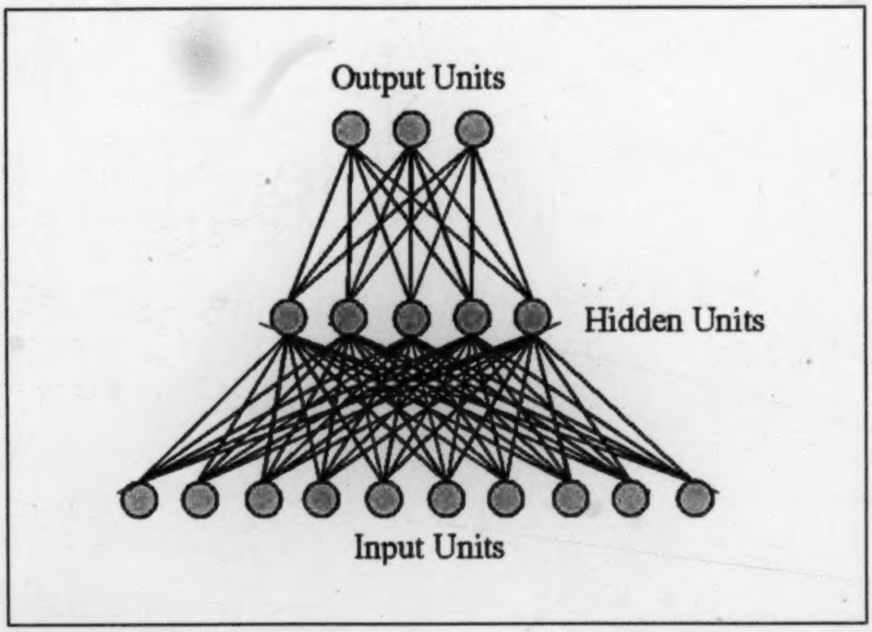

Figure 3-5. A multilayer perceptron.

Trained networks discover a mapping from the input space to the output space: they are trained to determine the associative relationship between the input patterns and their class (or their expected output value). The hidden layer of units in multi-layer perceptrons provides the network an opportunity to restructure the data into sub-groupings of its own choosing before it computes an output. Networks with connections only between an input layer and an output layer rely on the representation and organization offered in the input space.

Associative Learning and Connectionist Learning Rules

Uses of connectionist architectures to demonstrate particular phenomena of learning are not unknown in the literature (e.g. Gluck & Bower, 1988, 1990; Gluck & Thompson, 1987; Kehoe, 1998; Pearce, 1994; Schmajuk, 1997; Shanks, 1991). Many notable examples, however, (e.g. Shanks, 1995) receive connectionism primarily as a technique useful for the elucidation of associative learning phenomena within traditional models. Some, however, have allowed for the possibility of collaborative work among fields within the cognitive areas. In their recent review of theories of associative learning, Wasserman and Miller (1997) welcome interdisciplinary efforts to understand learning:

> In accord with Thorndike's and Pavlov's early speculations, elementary associative learning still seems able to serve as the foundation for our understanding of many complex forms of behavior and cognition. However, our review reveals a rich body of knowledge about associations that surely causes us to question the simplicity of even this basic brand of mentation . . . The next several years of research will be exciting ones, as neuroscientists and cognitive

> scientists join experimental psychologists in an interdisciplinary attack on the challenging problems of associative learning and behavior change. (p. 598).

In this section, I argue that further integrations and collaborations between connectionist and learning theorists are essential to the progress of these sub-fields within their disciplines. Connectionist models can bring something to a union between these two areas that learning theorists have long acknowledged as a shortfall of traditional learning theories (e.g. Spence, 1936) – representational power. Where connectionist networks have floundered, however, learning theory is strong – in the acknowledgment of specific processes of learning that many cognitive science models have overlooked. According to Hanson & Burr, connectionism "is especially suited to learning and allows the relationship between learning and representation to be studied directly for the first time"(1990, p. 471). Given this, connectionist models provide an interesting and valuable common language to learning theory and cognitive conceptions of cognition.

As mentioned earlier, neural networks learn. The process of learning is through modification of the connections between units in the network. This modification is done according to a *learning rule*. One powerful learning rule is the Widrow-Hoff rule (Widrow & Hoff, 1960/1988; or the Delta rule, Rumelhart and McClelland, 1986) – a least mean squared method for adjusting the network weights such that the difference between the desired output and the actual output is minimized. The Widrow-Hoff rule is used to train simple connectionist networks like perceptrons, in which processing units from the input layer are connected directly to the units in the output layer, with no layer of intervening hidden units.

In 1981, in an introduction of their model of learning, Sutton and Barto showed the equivalence of the Rescorla-Wagner model of learning and the Widrow-Hoff learning rule, given certain constraints. This proof demonstrated the close relationship that exists between traditional learning theories and models of learning that have developed in cognitive science.

Although Widrow-Hoff is a powerful learning rule, it can be demonstrated that the networks trained with this learning rule are not representationally powerful enough to account for learning in *all* situations (Minsky & Papert, 1969/1988). As an example, networks trained with the Widrow-Hoff learning rule can only learn to classify a set of patterns if the set of input vectors meets the constraint of *linear separability* (Bechtel & Abrahamsen, 1991). An "exclusive OR" problem (X-OR; referred to in learning research as a negative patterning problem) does not meet the constraint of linear separability, as the learner must learn to discriminate on more than one dimension, and must carve the problem space in a way that is non-linear; therefore, the Widrow-Hoff rule is representationally inadequate to account for X-OR (Dawson, 1998) or negative patterning learning, as is suggested in the following chapter.

As a model of learning, the Widrow-Hoff rule lacks the power to train networks that are able to represent certain types of associative learning; its learning-theory equivalent, the Rescorla-Wagner model of learning can, likewise, be shown to be limited. Some solutions to the representation problem have been attempted in learning theory (Bitterman, 1953; Pearce, 1987; 1994). These solutions require the assumption that the learning organism treats some element of a compound stimulus as different, either by hypothesizing a separate unique cue that is associated with the compound stimulus, or by assuming that the compound stimulus is treated holistically, but differently than the elements that make up the compound. The assumption that a compound stimulus is represented separately from the elements of the compound is explored in greater detail in Chapter 4 in two connectionist architectures.

In cognitive science circles, the criticism of early connectionist models trained with the Widrow-Hoff learning rule stalled connectionism in the 1970's and necessitated a major change in connectionist models. Any model of learning must be powerful enough to solve linearly non-separable problems since these problems are routinely solved by organisms in everyday life. In fact, humans can learn linearly non-separable problems at least as fast as linearly separably ones (Medin & Schwanenflugel, 1981). The Widrow-Hoff rule is limited in a way that learning in nature is not and is therefore not a plausible model of learning.

The problems associated with the failures of the Widrow-Hoff rule were addressed by building multi-layer networks and training them using a learning rule called *back-propagation* (also called the generalized delta rule; Rumelhart, *et al.*, 1986). Back-propagation, like the Widrow-Hoff rule, uses the difference between the desired and actual output of the network to modify weights. The Widrow-Hoff rule is inappropriate for training multi-layer networks because it assumes that each modifiable connection weight is adjacent to an output unit. In multi-layer networks, the connections feeding the hidden units have no adjacent output units. Back-propagation is so called because it *propagates* the error from the output unit *back* through the network.

Back-propagation is an improvement over Widrow-Hoff learning because, theoretically, a multi-layer network trained using the back-propagation of error learning rule can represent *any* data set. Therefore, these networks are able to represent problems that are not linearly-separable like X-OR or negative patterning. According to Sarle, multi-layer perceptrons are "general-purpose, flexible, nonlinear models that, given enough hidden neurons and enough data, can approximate virtually any function to any desired degree of accuracy. In other words, MLPs are *universal approximators*" (1994, p. 5). According to Rumelhart, *et al*.: "if we have the right connections from the input units to a large enough set of hidden units, we can always find a representation that will perform any mapping from input to output through these hidden units" (1986, p. 319).

Connectionism and the Tri-level Hypothesis

In 1995, learning researcher David Shanks summarized the current state of the area of associative learning in *The Psychology of Associative Learning* and devoted a large section of his book to connectionism. As mentioned above, connectionist architectures have been used by researchers as a method to elucidate particular phenomena. Shanks book seems to mark the beginning of a new era in learning – an era in which connectionism is accepted as a *mainstream* associative method. Shanks treatment of connectionism provides a link between connectionist methods and learning theory. He uses the Tri-level hypothesis of David Marr (1982) as a framework for his argument supporting the use of connectionist architectures for modeling learning.

Marr's theory holds that the nature of cognitive science in general and information processing in particular is such that multiple levels of explanation are required to achieve a full understanding of an information processing system and, by extension, cognitive phenomena. Marr's three levels of explanation are the computational level, the representational/algorithmic level, and the implementational level (see also Dawson, 1998; Pylyshyn, 1984).

The *computational level* of analysis is concerned with the goal of the computation and, therefore, about defining the problem that the information-processing system is seeking to solve. Computational approaches involve translations into a formal language. Translations allow the development of strategies for solving problems and make predictions about the system (Dawson, 1998). The *algorithmic level* deals with process questions – "how is the goal of the system realized?" or "what information processing steps are used to arrive at the solution to the problem being solved?". The *implementational level* asks questions about the physical components of the system - "how are the information processing steps of the algorithmic level physically implemented in the system?". A complete explanation of cognitive processes requires answers to questions at each of these three levels of analysis.

Shanks answers the first question, the computational level question, with contingency theory: The system "computes the degree of conditional contingency between events" (1995, p. 104). He refers particularly to three different formal contingency-based models. These are χ^2, ΔP, and ΔD. These have been described in more detail in Chapter 2.

Shanks answers the second question with instance theories, and in particular, the context model (Medin, 1975) that emphasizes "the memorization of instances, with stimuli being represented in a multidimensional psychological space and with inter-stimulus similarities being an exponential function of distance in the space" (Shanks, 1995, p. 104). For context theories, each new instance is represented in memory in the category that contains the instance to which it is most similar.

The third question, that of implementation or mechanism, Shanks answers with connectionism. He identifies some of the limits of old connectionist architectures and their similarity to the Rescorla-Wagner model of learning. He argues that the additional power of multilayer connectionist networks to classify problems such as the X-OR problem makes them appealing as implementations of contingency theory. He describes the relationship between contingency and connectionism in terms of the mathematical equivalence, given certain constraints, of the delta rule (the Widrow-Hoff rule described earlier) and the Rescorla-Wagner model. He goes on to cite a proof by Chapman and Robbins (1990) that, in a network described with only input and output units and trained with the delta rule, the weight associated with a particular cue (input) will equal ΔP – "the degree of statistical contingency between the cue and the outcome" (Shanks, 1995, p. 114). This relationship holds only at asymptote. Shanks also describes a study by Wasserman, Elek, Chatlosh, & Baker (1993) in which Wasserman and colleagues asked human subjects to estimate the contingency between a key press and a flashing light for a number of different levels of P(O/A) and P(O/-A). The subjects' estimates were a very close fit to predictions made by the delta rule, at asymptote. Prior to asymptote, the delta rule provides a relatively good fit to acquisition curves (Shanks, 1995, p. 114).

Shanks' attempt to harmonize various theories of learning by assigning them to different levels of the tri-level hypothesis is an interesting proposal. However, when the three levels of explanation are introduced by Marr (1982) as the appropriate method for describing cognitive phenomena, Marr also argues that these levels must be logically related to one another (see also Dawson, 1998; Pylyshyn, 1984; Rumelhart & McClelland, 1985; 1986).

If Shanks proposes contingency theory, in particular ΔP or ΔD, as the computational explanation of choice, he ought also to acknowledge that the appropriate, logically related implementation of contingency theory is an old connectionist network like a perceptron, trained with the Widrow-Hoff learning rule. Contrary to Shanks argument, multi-layer connectionist models trained using back-propagation of error do not extend the representational power of contingency theory. Rather, these models serve to highlight the representational inadequacy of contingency theory. For Shanks' levels of explanation to be theoretically consistent and logically related, he either needs to relinquish traditional conceptions of contingency learning at the computational level, or abandon the additional representational power that modern connectionist models provide.

Where, then, does connectionist theory fit in Marr's framework? Connectionism provides some answers to some questions at each of Marr's levels of analysis – the implementational, the algorithmic, and the computational.

Connectionism: an Implementational Account

Connectionism could be an interesting implementational account of cognition. However, connectionist models are not actually meant to model the physical processes at work in the brain in terms of containing one-to-one relationships of neurons to processing units – at least, not yet.

Connectionist models are information processing systems that are "brain-like" in structure. According to Bechtel, "PDP models are not themselves neural models – they are abstract processing schemes built on analogy with neuronal nets but capable of being realized in other architectures" (1985, p. 54). These networks are "neuronally inspired" (Rumelhart & McClelland, 1986, p. 130) as many of their processes are functionally similar to neural processes but they are not meant to be biologically equivalent to the brain or even to parts of the brain. In particular, the learning rule generally used to train multilayer perceptrons, back-propagation, has been criticized for its neural implausibility: "back-propagation is biologically implausible, inasmuch as error signals cannot literally be propagated back down the very same axon the signal came up" (Churchland & Sejnowski, 1989). Rumelhart and McClelland (1986) claim that they have made some decisions, for practical reasons, that simplify the process of cognition. Through this simplification, some biological plausibility is lost. They justify this by seeing the process of "model building as one of successive approximations" (Rumelhart & McClelland, p. 136).

Their functional equivalence is debated in the literature and attempts have been made to introduce connectionist models that are more "biologically plausible" (Gluck & Myers, 1993; 2001; Schmajuk, 1997). It is generally agreed, however, that despite the numerous variables that make a general, biologically equivalent model unlikely at this time, PDP models are biologically and functionally *similar* to brains. This makes these models potentially more interesting than their cognitive psychology, flow-chart competitors or than their behavioural psychology ΔP or ΔD ancestors.

While connectionist models are, to some degree, implementational accounts, in that they are functionally similar to the physical structure they model, they can also be considered at Marr's second level of analysis, the algorithmic level.

Connectionism: an Algorithmic Level Account

Algorithmic explanations are concerned with algorithms followed or steps taken in information processing. It is this level that has been the traditional focus of cognitive and behavioural psychology. Most research and theorizing in these two sub-specialties of psychology is done on psychological processes (algorithms), not at either the formal language (computational) level or the biological implementational level. While connectionist conceptions of cognition have biological themes and can speak to issues of implementation, and can have formal aspects and can speak to issues of computation (as we will see in the following section), most connectionist models can be understood best as theories spoken in the algorithmic language, as models of *processes* of cognition. According to Rumelhart and McClelland, "we believe that we are studying the *mechanisms* of cognition"(1986, p. 120).

In a discussion of recent evidence and the plausibility of principles of association, Rescorla (1992) suggests that CS-US associations may need to be considered *hierarchically* to account for the data. He further suggests that hierarchical associative functions are implemented in a multilayer network, aligning the notions of the hierarchical associative model and connectionist architectures. He describes the primary level of explanation at which this argument is made for instrumental learning: "Furthering the understanding of how hierarchical organization is learned and functions is a prime issue in the analysis of this sort of learning" (p. 70).

While the algorithmic level may be the primary level at which PDP models have been considered by some of their proponents (Rumelhart & McClelland, 1986), computational level questions can also be answered by connectionist theory.

Connectionism: a Computational Level Account

There are an almost infinite number of combinations of components of PDP networks, therefore, a simple, formal explanation of what PDP networks are computing in general is impossible. Most PDP networks, however, are computing one of a number of complex forms of multiple regression. According to Sarle (1994, p. 1), "multilayer perceptrons are nothing more than nonlinear regression and discriminant models." As networks become more complex, however, so do the statistical algorithms they resemble. Some models, such as counterpropagation and self organizing maps (Sarle, 1994) are best defined as nonstatistical as they have no statistical equivalents. A perceptron with a linear activation function is similar to multiple or multivariate linear regression (depending upon the number of output units): if the activation function is logistic, the analogous statistic is logistic regression. When nonlinear hidden units are added to perceptrons, the math becomes more intricate but the algorithm is still similar to various forms of nonlinear multiple regression.

Connectionism as Cognitive Theory

Marr's three levels of analysis must be logically related: they also act to constrain one another. Implementational accounts are interested in how cognition might actually be effected in the brain. Implementational accounts are related to algorithmic accounts in terms of how the information processing steps might be represented. As an example of how one level may constrain the theorizing that is done at a different level, consider the following: "any algorithm that would require more specific events to be stored separately than there are synapses in the brain should be given a lower plausibility rating than those that require much less storage" (Rumelhart & McClelland, 1985, p. 194). Storage is really about representation, an issue primarily considered at the algorithmic level of analysis. Representation is relevant at the computational level as well, although the method of representation is not important. Computational level analysis of representation only require that the "representation is rich enough, in principle, to support computation of the required function" (Rumelhart & McClelland, 1985, p. 194) and that the algorithmic level explanation of representation is consistent with the computational level.

Implicit in the primarily algorithmic account of cognition and learning that connectionist models provide is an implementational account and a computational account (Rumelhart & McClelland, 1985). The logical, computational story associated with connectionist models at either the algorithmic or the implementational levels is not contingency. From that perspective, connectionism must be considered a competitor of simple associative theory. However, " . . . any behavior that can be characterized by associative principles can *ipso facto* be characterized by the more powerful models. Such models should not, therefore, be considered as alternatives to associative models; rather, associative rules are simply special cases of the rules employed by more powerful theories" (Bever, Fodor, & Garrett, 1968, p.585).

Connectionist models can produce answers to questions at each of Marr's three levels of analysis. The answers that are provided by these models are logically and causally related. Is this enough evidence to consider connectionism a theory of cognition? This question is one that is debated in the literature. Despite the fact that it is unlikely a complete cognitive account in its present form, PDP theory provides an appropriate approximation to cognition.

Considering connectionist models as theories is not new and has been controversial. McCloskey (1991), for example, suggests that there is a large gap between simulation and explanation. He describes a scenario in which a black box is connected to a keyboard and a monitor. Input in the form of a letter string is given through the keyboard and a word/non-word decision is printed to the screen. A phonological representation is printed to the screen. A reaction time for each of these processes is also displayed. He then suggests that it would be possible to test this machine in a number of different situations and establish relatively good correspondence between the responses of the machine and the responses of a human sample. From here, McCloskey asks whether his black box is a theory of word recognition. It impressively classifies word/non-words similarly to human samples. However, in order to classify such a machine as a theory, you would want to ascertain how it arrived at a lexical decision; how it represented phonology; how particular phenomena of interest came to be seen in machine's functioning. Even if all the components of the black box are described in detail (McCloskey does this and, in this particular thought experiment, the components are connectionist after Seidenberg and McClelland, 1989), McCloskey suggests that there is insufficient information provided to allow the machine to be considered a theory. He argues that "although the ability of a connectionist network (or other computational device) to reproduce certain aspects of human performance is interesting and impressive, this ability alone does not qualify the network as a theory, and does not amount to explaining performance" (1991, p. 388).

The Connectionist Black Box

McCloskey's criticism of connectionism is another echo of a theme in behaviorist psychology: the problem of the "black box". Behaviourist theories have been criticized as being unfalsifiable: "black-box theories have an explanatory ceiling that cannot be penetrated because of an intrinsic inability of black-box theorists to control the causes of behavior" (Kendler, 1989, p. 265). Both connectionists and behaviourists have been reproached for being too concerned with *what* organisms are doing and not concerned enough with *how* the organisms are doing what they are doing. However, most connectionists actually consider PDP models primarily as algorithmic accounts of cognition, as mentioned previously (Rumelhart & McClelland, 1986), and have been very interested in representations (Hanson and Burr, 1990; Maki, 1990). Rumelhart and McClelland suggest that this is the primary difference between behaviourist accounts and connectionist accounts: "In our models, we are explicitly concerned with the problem of internal representation and mental processing, whereas the radical behaviorist explicitly denies the scientific utility and even the validity of the consideration of these constructs" (Rumelhart & McClelland, 1986, p. 121).

These representations, however, rarely pop out of connectionist models: the things that make PDP models interesting (parallel processing; distributed representations) also make them complicated. Although complete explanations of representations may be desirable for connectionist researchers, representations are often obscured in models, making explanations difficult. Connectionist models have been accused of being entangled in *Bonini's Paradox*: "If a computer simulation falls into this trap, then this means that it is no easier (and perhaps is harder) to understand than the phenomenon that the simulation was suppose to illuminate" (Dawson, 1998, p. 123). PDP models can easily become very complicated and have, most often, been considered impossible to interpret. According to Seidenberg, "Connectionist models do not clarify theoretical ideas, they obscure them" (1993, p. 229).

If connectionism is to be considered a tool of clarification, connectionist researchers must move toward a policy of looking into the "black box" to determine how their models are forming and storing representations. This is required if connectionism is to be considered an algorithmic account of cognition: process and representation are key to this level of analysis. Interpretation of PDP models may, then, be essential if we hope to use connectionist models to inform psychology about theory, which is how McCloskey (1991) concludes his argument. According to Rumelhart and Todd "getting a coherent picture of 'what goes on' inside a network as it develops, manipulates, and alters the representation of the knowledge it processes is vital for our understanding of connectionist information processing, and likely for our understanding of the minds these systems model" (1992, p. 3). If connectionist models can be shown to have sensible internal structure, they may be considered not only *interesting*, as is McCloskey's (1991) weak compliment to connectionist models, but also *informative*.

Some researchers have made attempts to understand the internal structure of networks, as McCloskey (1991) recommends as a solution to the black box. Some are interpreting their networks in terms of regularities in activations of hidden units (e.g., see Berkeley, Dawson, Medler, Schopflocher, & Hornsby, 1995; Christiansen & Chater, 1992; Dawson, Medler & Berkeley, 1997; Elman, 1990; Gorman & Sejnowski, 1988; Hanson & Burr, 1990; Willson, Valsangkar-Smyth, McCaughan, & Dawson, 1999) and some have purposely inserted structure into their networks (e.g. McMillan, Mozer & Smolensky, 1991; Dawson, Medler, McCaughan, Willson, & Carbonaro, 2000).

Connectionist models can, given certain conditions, be considered theories of cognition. Given that these models can also provide answers at each of Marr's levels of analysis, the implementational, the algorithmic, and the computational, with future "successive approximations" (Rumelhart & McClelland, 1986, p. 136) they may one day be considered complete theories of cognition and learning. For now, it may be enough to say that successful aspects of these models can be used to direct cognitive theory. Hanson suggests that "the simple assumptions in connectionist models that lead to successful practical results can generate theories" (1990, p. 511).

General Discussion

The relationship between connectionism and associative learning theory may not be as Shanks describes – with connectionism as an implementational account of contingency – but it does exist. Connectionism is associationist, but not merely associationist, according to Bechtel and Abrahamsen (1991). Connectionism is:

> an elaboration of associationism that has benefited (*sic*) from and can contribute to many of the goals of the cognitivism of the last twenty years . . . Among the elaborations that were not even conceived of within classical associationism are: distributed representation . . . hidden units . . . mathematical models of the dynamics of associationist learning, supervised learning . . . back-propagation, and simulated annealing within a self-organizing dynamic network" (p. 102).

If these properties of networks give multilayer networks the edge over perceptrons, their presence in models of associative learning may provide those models with the power to solve the kinds of problems that multilayer networks are able to solve.

Conceptions of cognition and learning are needed, that adopt features that are known to have computational power, algorithmic validity, and implementational plausibility. These properties are available in connectionist models. However, while learning rules most commonly used to train connectionist architectures may be representationally powerful and implementationally interesting, they have been criticized for failing to be sensitive to known principles of learning – an algorithmic level criticism.

As an example, consider the concept of the back-propagation of error through a network. The basic theorem dictates that feedback is given to a network that is making an incorrect classification about *how wrong* its classification is. In most learning situations, this type of feedback is not provided. Rather, an organism is generally limited to information about whether a particular behaviour is right (through the presence of a reinforcer) or wrong (through the absence of a reinforcer).

Some attempts to rectify this limitation have been made, most notably by Sutton

& Barto (1998) who have introduced "reinforcement learning" to artificial models. Reinforcement learning is an approach in which only global feedback is given to the network. Rather than supervising a network and providing the differences between actual and expected responses, reinforcement learning only provides information about whether a response is right or wrong. This may be a more realistic method for training networks, however, reinforcement learning is cumbersome and training times are often excessively long (Hinton, 1989).

Another weakness of back-propagation that has received some attention in the literature is a problem known as *catastrophic forgetting* (Ratcliff, 1990; Robins, 1995). Catastrophic forgetting occurs when networks trained in a situation *a* are retrained in a situation, *b*, then retested on problems in situation *a*. Networks trained using back-propagation tend to have poor recovery of learning in situation *a*; the learning that has taken place in situation *b* writes over the prior learning. The phenomenon of catastrophic forgetting in neural networks makes these networks dubious models of learning, since a learning organism is easily able to overcome the challenge of learning associations in new situations, then needing to recall associations in prior situations. This phenomenon will be dealt with in greater detail in Chapter 5.

Recently, Delamater, Sosa and Katz (1999) have demonstrated a use of a connectionist model in learning theory construction that goes beyond the conception of connectionism as an implementational tool. They use a connectionist network, trained using back propagation, to suggest hypotheses for a study that was designed to evaluate the two main accounts of discrimination learning: the unique cue hypothesis, and the configural approach. They then use the structure of the network to conclude that contributions to discrimination learning may be made both by unique cues and configural cues. The Delamater *et al.* study illustrates the use of a connectionist model to guide theory as applied to a specific learning task: discrimination learning. I contend that connectionist models can be exploited further – as will be explored in the following chapter.

One of the properties of connectionist models considered by Bechtel and Abrahamsen (1991, p. 102) to be extensions of classical associationism that ought to be considered in modern cognitive models is the notion of distributed representations. This issue is considered in detail in the following chapter, in the context of discrimination learning.

Chapter 4

THE "PROBLEM OF PATTERNING" REVISITED

Over the years, patterning has generated some controversy in the field of discrimination learning because it has been difficult to explain in terms of accepted theories of representation. *Patterning* is a special case of conditional discrimination. *Positive patterning* occurs when a subject is trained to respond to a compound stimulus, AB, but not when either A or B is presented alone. *Negative patterning* occurs when a subject responds to either A or B alone, but not when A and B are presented in compound.

Negative patterning is a problem for associative theory because both stimulus A and stimulus B are presented as often without reinforcement as they are with reinforcement. There should accrue neither a positive nor a negative association between the stimuli and the outcome; in a balanced study, both stimuli have a 50 percent chance of occurring with reinforcement. This is a problem for one of the early assumptions of conditioning: that associative strength conforms to an additive principle such that the associative strength of a compound stimulus, AB, is equal to the sum of the associative strengths of the components, stimuli A and B (Spence, 1936). The summative principle is inadequate to account for negative patterning because subjects are taught that stimulus A alone has a positive weight, stimulus B alone has a positive weight, but that the compound stimulus AB has a negative weight. The summative principle, in this case, predicts a strong approach response to the compound AB, and cannot account for the avoidance learning.

This limitation of the summative principle parallels the issue in connectionism in which perceptron-like networks were shown to be unable to learn tasks that are linearly non-separable (Minsky & Papert, 1969/88). Negative patterning, described above, is formally equivalent to the linearly non-separable logic problem X-OR, described in Chapter 3; therefore, the summative principle and perceptrons are similarly limited. In connectionist networks and in models of patterning learning, this limitation can be overcome in more than one way.

One way of overcoming this problem in perceptrons is by transforming the input space – adding an extra input to a X-OR problem that signals the co-occurrence of the other input units modifies the problem such that it can be performed by a simple perceptron (Rumelhart, *et al.*, 1986).

A second way of overcoming this limitation in perceptrons is through the addition of hidden units that pre-process information. This approach has been considered in Chapter 3. In the hidden layer, a hidden processing unit can be used by the network to detect the co-occurrence of the input units. This enables the network to treat the compound stimulus differently than the elements that compose the compound, and solve the problem.

In models of discrimination learning there is also more than one way to solve the "problem of patterning" (Bitterman, 1953, p. 123). The solutions that have been offered can be classed into two groups. The first group modify the summative principle by the addition of a "unique cue" – this is parallel to the neural network solution of transforming the input space by the addition of an extra input unit. The second group of solutions overcome the limitation by assuming a process in which the co-occurrence of the inputs creates a configural representation that is treated holistically and differently than the elements that make up the compound – this is parallel to the neural network assumption of preprocessing hidden units. These learning solutions will be considered in the next sections.

Solutions to the Problem of Patterning

Patterning: A Unique Cue Account

To account for the limitation of the summative principle, alternatives were introduced, the earliest of which was the unique cue approach (Bitterman, 1953). According to this theory, a subject distinguishes between a compound stimulus and the components of that stimulus via a "unique cue" that is present during the compound stimulus trials but not present on the simple stimulus trials. In the negative patterning situation, A and B both acquire a positive weight but when they are presented together, the unique cue, U_{AB}, has a strong enough negative weight that it overcomes the sum of the positive weights of A and B.

Improvements to the simplest unique cue model include the influential Rescorla-Wagner model (Rescorla & Wagner, 1972) that, while it explains some phenomena that are problematic for the simple unique cue approach, fails to explain other phenomena such as single trial conditioning and some types of discrimination problems (Pearce, 1994).

In general, these theories can be referred to as elemental theories: they depend upon representations of the elements and the summation of the associative strengths associated with these elements to account for compound conditioning.

Patterning: A Configural Cue Account

An alternative to unique cue approaches has been the configural cue account, in which a compound stimulus AB is treated holistically, and differently than the elements that make up the compound (Spence, 1952). The representation of the compound stimulus is a single representation but it is affected by the similarity of the compound to other stimuli (Gluck, 1991).

In support of a configural hypothesis, Pearce (1994) has developed a connectionist model with a layer of input units, a layer of what Pearce calls "output units" that intervene between the units in the configural layer, a layer of hidden "configural" units and an output unit that Pearce identifies as an unconditioned stimulus. There are as many "output units" in the model as there are elements and as many configural units in the model as there are combinations of these elements. Each of the configural units is prepared to key in on a particular stimulus; either an elemental stimulus or a compound stimulus (e.g. A or AB or ABC *etc.*). Connections between units remain latent until explicit training occurs that awakens particular connections in the network. Pearce's model has had success in predicting many learning phenomena and has been very influential in extending theories about compound stimuli.

Elemental and Configural Processes

Some researchers who have been concerned with the learning of compound stimulus tasks have begun to look at how these processes may interact and whether other processes may be at work. Kehoe and colleagues, for example, (Bellingham, Gillette-Bellingham, & Kehoe, 1985; Kehoe, 1986; 1988; Kehoe & Gormezano, 1980; Kehoe & Graham, 1988; Weidemann & Kehoe, 1997) have investigated negative and positive patterning, primarily in the nictitating membrane response of the rabbit. In this particular preparation, Kehoe and colleagues have observed both summative processes and configural or "Gestalt-like" processes. In a study evaluating theories of the conditioning of compound stimuli, Bellingham *et al.* (1985) found that a simulation of the unique cue hypothesis predicted that positive patterning should proceed more slowly than negative patterning: a prediction that is contradicted by the empirical data presented. They also simulated the configural hypothesis in which there was some generalization between stimulus elements and compounds to which the elements belonged. They found that the configural hypothesis failed to predict excitatory summation early in training during acquisition of the negative patterning task. Kehoe and Graham (1988) in a study investigating stimulus compounding (training with two distinct stimuli) and negative patterning, found no support for a pure configural hypothesis and some support for a unique stimulus hypothesis.

Taking this mixed evidence into consideration, it seems unlikely that either a simple unique stimulus theory or a purely configural theory is able to account for the complex process of learning compound stimulus tasks. While an elemental theory may be better able to account for stimulus generalization and initial response to the compound in negative patterning, configuration describes a more parsimonious, although still somewhat incomplete, solution to the negative patterning task.

Recently, Delamater and colleagues (1999) have restated that Pearce's configural model is inadequate to account for some phenomena. They have proposed that a combination of Pearce's configural account and an elemental account better predicts experimental results. The Delamater *et al.* model looks similar to the Pearce model: the same number of input units and configural units, and a single US output unit. The primary difference between the models is in the connections between the units. While Pearce's model contains only connections between units in one layer and units in the next layer, Delamater *et al.* have added connections between the input layer and the US output unit that skip the intervening layer. Recall from Chapter 3 that two-layer networks – networks without any hidden units – that are trained with the Widrow-Hoff learning rule are instantiations of the Rescorla-Wagner Model (Rescorla and Wagner, 1972). The Rescorla-Wagner model is a variant of the simple elemental hypothesis: while the connections in the Pearce model are "configural connections", the connections linking the input layer directly with the output layer add a "unique cue" or Rescorla-Wagner component to the Delamater *et al.* model.

Linear Regression Model

In statistical terms, the unique cue model, the configural account and the blend of the two, can all be viewed as linear regression models in which A and B are predictors, and either U_{AB} or AB or the configural unit that responds to input AB represents the interaction between them. In a problem such as the negative patterning problem described above, in which there are four states (A off, B off; A on, B off; A off, B on; A on, B on), there are five terms required in the regression equation $(y_{ijk} = \mu_T + \alpha_j + \beta_k + \alpha\beta_{jk} + \varepsilon_{ijk})$.[2] In a slightly more complex problem in which there are three stimuli (A, B, and C), there are eight (2^3) possible states and nine terms in the regression equation $(y_{ijkl} = \mu_T + \alpha_j + \beta_k + \lambda_m + \alpha\beta_{jk} + \alpha\lambda_{jm} + \beta\lambda_{km} + \alpha\beta\lambda_{jkm} + \varepsilon_{ijk})$.[3] A subject learning a negative patterning problem with just three inputs would be required to learn and remember eight contingencies. If the problem were increased to seven stimuli, the subject would need to learn and remember 128 (2^7) separate contingencies – a difficult problem.

In a discussion about representations in models of learning, Kehoe (1988) acknowledges the special challenge to models of discrimination learning posed by negative patterning. He discusses the process of converting a non-linear problem like the simple, two element negative patterning problem into a linear problem with the addition of a special input that responds to the joint occurrence of the two elements. This move is one that is equivalent to the move made by the elemental theory camp with the introduction of the unique cue. According to Kehoe, however, "this tactic for solving nonlinear representation problems would create an explosive proliferation of special inputs" (1988, p. 412; also Kehoe, 1990).

---

[2] The general linear model in which y_{ijk} is the score of the i_{th} subject in the j_{th} level of A_j and the k_{th} level of B_k; μ_T is the grand mean; α_j is the treatment effect of factor A_j $(\mu_{Aj} - \mu_T)$; β_k is the treatment effect of factor B_k; $\alpha\beta_{jk}$ is the interaction effect of treatments A_j and B_k ; and ε_{ijk} is the error component of the equation.

[3] As above, the general linear model in which y_{ijkl} is the score of the i_{th} subject in the j_{th} level of A_j, the k_{th} level of B_k and the m_{th} level of C_m; μ_T is the grand mean; α_j and β_k are as above; λ_m is the treatment effect of factor C_m; $\alpha\beta_{jk}$, $\alpha\lambda_{jm}$ and $\beta\lambda_{km}$ are 2-way interaction effects; $\alpha\beta\lambda_{jkm}$ is the 3-way interaction effect of treatments A_j, B_k, and C_m; and ε_{ijk} is the error component of the equation.

Whether the special units are input units in a two-layer network or are extra hidden units in a multi-layer network that are required to provide representations of the compound stimuli (as in the configural cue account of compound stimulus learning), it seems evident that the number of processing units required is a power function of *n*, in which *n* is the number of elements in a discrimination problem.

Other connectionist models of configural learning that have demonstrated some success in accounting for phenomena associated with compound stimuli can also be critiqued on this level. Schmajuk and DiCarlo (1992), for example, whose model inspired the unique cue connections in the Delamater *et al.* network, present a connectionist model in which two simple stimuli are presented as inputs along with an input unit representing the context. These three inputs are massively connected to three configural stimulus nodes that are massively connected to six "simple and configural stimulus-US associations". These configural and simple association nodes are also massively connected to the 3 nodes in the input layer. The 6 configural and simple association nodes are connected to a single output unit, the US. Schmajuk and DiCarlo's model has had great success at accounting for many discrimination learning phenomena (for a list of the model's successes, see Schmajuk, 1997). However, the number of parameters in this model is large for the simple tasks it is designed to perform: the discrimination of 2 elements and their compound. This type of solution to the "data fitting" problem in discrimination learning is a likely case of overfitting, similar to a regression equation with many more predictors than data points to predict.

A promising and manageably-sized network model of discrimination learning has been proposed by Kehoe (1988). Kehoe's model has successfully predicted many learning phenomena. In this model, two elements T and L are presented as sensory inputs linked to two hidden units that are connected to a single output unit. The model also contains a single US sensory input unit that is connected both to the hidden layer and directly to the output unit. This US unit, in spite of its name, actually has the ability to function as a special input in the simple discrimination tasks that this network was designed to perform. Whether this particular model would require additional "US" input units when scaled-up remains to be seen.

The "explosive proliferation" of parameters in recent models of discrimination learning requires us to consider that current conceptions of discrimination learning may be less than parsimonious. To further this argument, let us consider that, while the processing of associations is done in the mind, ultimately, representation is traced back to the physical structure, the *brain*, and the brain is a finite machine. Demonstrating that the brain does not have the capacity to solve simple discrimination problems by representing compound stimuli distinctly from their components can be considered a challenge to the unique cue hypothesis, the configural cue hypothesis, and any model that blends these but requires a configural or a special unit for each compound stimulus.

Ballard's Packing Constraint

The number of neurons in the brain is vast – but not infinite. The limitations of the brain have been informally recognized, but the concept of the boundedness of the brain was initially operationalized by cognitive scientist, Dana Ballard (1986). As an illustration, Ballard used a color discrimination task to demonstrate that neuronal demand can easily surpass supply when the assumption is made that neurons represent specific stimuli. This problem, termed the packing problem, can be quantified by the equation $U = N^k$ in which U is the number of processing units required, N is the number of inputs (in this case, number of discernable differences within each color category) and K is the dimensionality of the unit (or the number of colors considered). In line with this equation, Ballard found that U easily surpasses the number of neurons available, in order to perform a particular, reasonably simple, colour discrimination task.

By extension, assuming one neuron per stimulus, unique cue or configural cue encoding by the processes involved in some of the connectionist models discussed in this chapter, would also often surpass the limits of an animal brain for complex tasks such as navigating through an environment, locating a distant food source, or hunting moving prey.

Solving the Packing Problem

In order to solve a complex problem with multiple stimuli, a learner takes shortcuts. It is well understood that the notion that one neuron could be responsible for a single representation is an oversimplification. This oversimplication can be thought of as a localist assumption. We know, however, that a single neuron can be partially responsible for many representations, in concert with many other neurons, and it is a pattern of neuronal activation that allows discrimination among these representations.

The packing problem was first introduced by Ballard in the context of connectionism. Power limitations of locally coded models was acknowledged by other proponents of PDP modeling as well (e.g. McClelland, 1986; Rumelhart & McClelland, 1986). In these circles, the packing problem was resolved by the addition of some properties of connectionist models that allow for patterns of activation across units in a network. For example, connectionist networks can be constructed such that individual "neurons" are used in multiple representations in a manner analogous to overlapping receptive fields in the visual system. According to Hinton, McClelland and Rumelhart (1986), "each entity is represented by a pattern of activity distributed over many computing elements, and each computing element is involved in representing many different entities" (p. 77).

In order to accomplish representations that are distributed across hidden units in a connectionist model, it is necessary to restrict the number of hidden units in the model such that the number of hidden units is fewer than the number of input combinations to be represented. In a situation then, in which there are three stimulus elements, A, B, and C, and all the possible compounds of these elements, the number of hidden units must be less than 9, the number of input combinations. If a model has as many hidden units as it has input patterns, a hidden unit can key in on only one input pattern. Distributed representations are not required of such a model.

In the case of the models used by both Pearce (1994) and Delamater *et al.*(1999) there are as many hidden units as there are input patterns. This does not appear to be a problem for the learning of the simple tasks that these networks simulate, but becomes a problem when these models are extended to account for learning in more complex, "real world" tasks because the power that is required to solve this type of problem exceeds that which is available to the model.

Few scientists today would argue that any representation could possibly be contained in a single neuron. Both the logic of the packing constraint and our current knowledge about the function and anatomy of the brain make this argument implausible. Neurons are massively interconnected and interdependent. The primary assumption, therefore, guiding the packing constraint (one stimulus, one neuron) is, then, obviously false. Rather than nullifying Ballard's argument, however, this fact underlines Ballard's message: something besides local coding is happening in the brain. If we assume that some set of neurons is responsible for a particular representation rather than a single neuron, the capacity of the brain to represent large numbers of stimuli is even more apparent, unless we assume overlapping representations. Overlapping representations are a "given" in terms of neurobiology. Why shouldn't we be using this evidence to build more functionally and psychologically plausible models of learning, that are more and more often instantiated in "neuronally inspired" (Rumelhart & McClelland, 1986, p. 130) connectionist networks?

The Inductive Approach

While connectionist models are often used as deductive tools to make predictions based on a particular theory that is instantiated in the network, networks can also be used inductively to guide theory.

One advantage of connectionist systems over classical, rule-based systems in cognitive science is that a network is not required to produce representations that make semantic sense (Rumelhart & Todd, 1993). Rather, a network dynamically develops a set of representations that best fits the task of mapping the inputs to the outputs. While networks in which representations are intentionally locally coded are not unknown in the literature (e.g. McClelland & Rumelhart, 1981), networks with distributed representations are generally considered more desirable and interesting (Rumelhart & Todd, 1993). Encouraging local representations by providing sufficient hidden units to achieve this discourages the "semantics-free" property that is appealing in connectionist networks. As each representation can be found in a single hidden unit, the unit acts as a rule that classifies the input as some category of output. A fully locally distributed network is simply an instantiation of a theory in which the rules of the theory are wired, in advance, into the network.

While there is nothing particularly *wrong* with instantiating a theory in a network to derive novel predictions in a deductive manner, it is not the only way that connectionist networks can be used. By turning the question asked around, and starting, inductively, with the data, it is possible that we may be able to answer different types of theoretical questions.

Pearce's connectionist model (1994) is, deliberately, an instantiation of Pearce's configural theory of compound conditioning (1987, 1994). The network of Delamater *et al.* (1999) contains components that correspond to configural units and direct connections that correspond to elemental associations that are, again, deliberately included in the network architecture.

Some interesting questions arise from this discussion. What happens in a network if, rather than *hardwiring* or *handwiring*[4] it to perform an input/output mapping according to either a configural, an elemental or a blended theory, elemental or configural processes are allowed to *emerge* in a network during training? And what happens in a network if it is constrained such that one-to-one relationships of input patterns to hidden units are disallowed?

In the following simulations, there are fewer hidden units in the hidden layer than there are input patterns to classify in the particular problems to be learned. Not only does this encourage the networks to find their own input/output mappings but it also is closer to a solution that can fit within the constraints imposed by the packing problem and takes advantage of the power of a PDP model by asking a simple network to perform a complicated task.

Study 1: Configural Cues in Patterning

This study demonstrates negative and positive patterning learning in a model similar to Pearce's configural cue model, but with fewer hidden units than inputs, requiring distributed representations. The method simulates that of Delamater *et al.* who pretrained rats and networks on a task, then trained them on either a negative or positive patterning task, using previously reinforced elements or previously not reinforced elements.

The rats demonstrated facilitation of learning for the negative patterning task when the elements had been previously reinforced. They also found slight facilitation for the previously reinforced elements group for the positive patterning task. Relevant to our discussion here, they also found initial excitatory summation in both previously reinforced and not reinforced conditions in the negative patterning task, that is, the rats responded initially more to the previously unseen compound stimulus AV than to either element A or V alone, in spite of the fact that the elements were reinforced while the compound was not. Initial excitatory summation is consistent with findings from other labs in both negative patterning (in which elements A and B are reinforced while the compound is presented during training without reinforcement) and in stimulus compounding (in which elements A and B are reinforced during training and the compound is tested subsequent to the training) and with other preparations (Bellingham, *et al.*, 1985; Couvillon & Bitterman, 1982; Kehoe, 1986; 1988; 1998; Kehoe & Gomezano, 1980; Kehoe & Graham, 1988; Kehoe, Horne, Horne & McCrae, 1994).

---

[4] Here I distinguish between "hardwiring" a network with fixed units that are non-adaptive and "handwiring" an adaptive network that is required to perform a task in a determined manner, simply as a function of its architecture.

In the Delamater *et al.* study, the configural cue network failed to produce the same results as the rats. In particular, although the researchers could find initial excitatory summation in the previously reinforced condition in the negative patterning task, they were unable to produce it in the not reinforced condition.

This study is meant to be a replication of the patterning study of Delamater *et al.* (1999); the major exceptions being that it employes a network in which the number of hidden units is less than the number of inputs to be discriminated, therefore requiring representations that are not simple one-to-one relationships between input patterns and hidden units, and the network is trained simultaneously on negative and positive patterning.

Method

Networks

All of the networks trained in this study had 10 input units, 4 hidden units, and a single output unit. All processing units had logistic activation functions. The units in the input layer were each connected to all hidden units; the hidden units were all connected to the output unit. There were no direct connections between the input units and the output unit. Prior to training, the connection weights between processing units were randomly started between -.5 and +.5. The network is depicted in Figure 4-1.

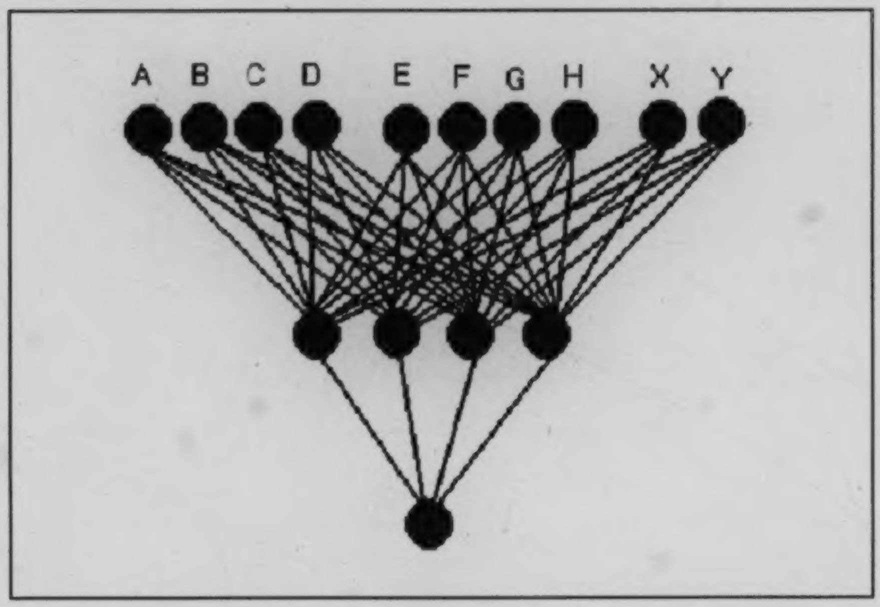

Figure 4-1. Simulation Study 1: A PDP model that simultaneously learns the negative and positive patterning problems. Units A - D are used to train the positive patterning task; Units E - H are used to train the negative patterning task; Units X and Y are background stimulus cues (see text for details).

Input Coding

Elements A through D were used for training on positive patterning, while elements E through H were used to train the network on the negative patterning task. Elements A, B, E and F were never presented together, and were always presented with background stimulus X , simulating similarity among these 4 stimuli (e.g. being auditory cues). Similarly, elements C, D, F and G were always presented with background stimulus Y, simulating a commonality (e.g. being visual cues).

In this simulated training situation, the input patterns were coded across 10 input units. Cues A through H corresponded to the first 8 (binary) input units of the network with a value of 1 indicating the presence of a particular cue and a value of 0 indicating its absence. Background cues X and Y were coded in the last two input units in the network and were, again, either off (0) or on (1). This design allows simultaneous training on both the negative and the positive patterning tasks, while providing a previously reinforced and a previously not reinforced condition for each task. The simultaneous training is undertaken in this study to underscore the representational power available in a relatively simple neural network.

Training

All networks in this study were trained using error back-propagation (Rumelhart, *et al.*, 1986) in which an error term derived from the difference between the expected output of the network and the actual output of the network is propagated back through the network and the connection weights are adjusted. Presentation of the input patterns was randomized. Connection weights were adjusted after each pattern presentation. The learning rate was 0.1 and all networks were pretrained and trained without momentum.

Pretraining. The pretraining phase consisted of presenting all networks with each element (A through H) with its appropriate background stimulus. AX, CY, EX, and GY were presented with reinforcement, while the remaining elements were presented without reinforcement. These pretraining input patterns were presented to the networks until the response of the network fell within .1 of the expected output state for all patterns.

Patterning. In the previously reinforced condition, the pretrained networks solved negative and positive patterning problems using previously reinforced elements (A, C, E, and G) in which the elements A and C were never reinforced but the compound always was reinforced (AX^0, CY^0, $AXCY^+$) and elements E and G were always reinforced but the compound never was (EX^+, GY^+, $EXGY^0$).

In the previously not reinforced condition, not reinforced elements were used (B, D, F, and H). B and D were used for the positive patterning task (BX^0, DY^0, $BXDY^+$). F and H were used for the negative patterning task (FX^+, HY^+, $FXHY^0$).

Networks were trained until their output fell within .1 of the expected output for all patterns in the training set.

Results and Discussion

As in the Delamater *et al.* (1999) study, these networks did not display initial excitatory summation in the previously not reinforced condition for the negative patterning task, when that task was isolated from the positive patterning task for analysis (see Figure 4-2).

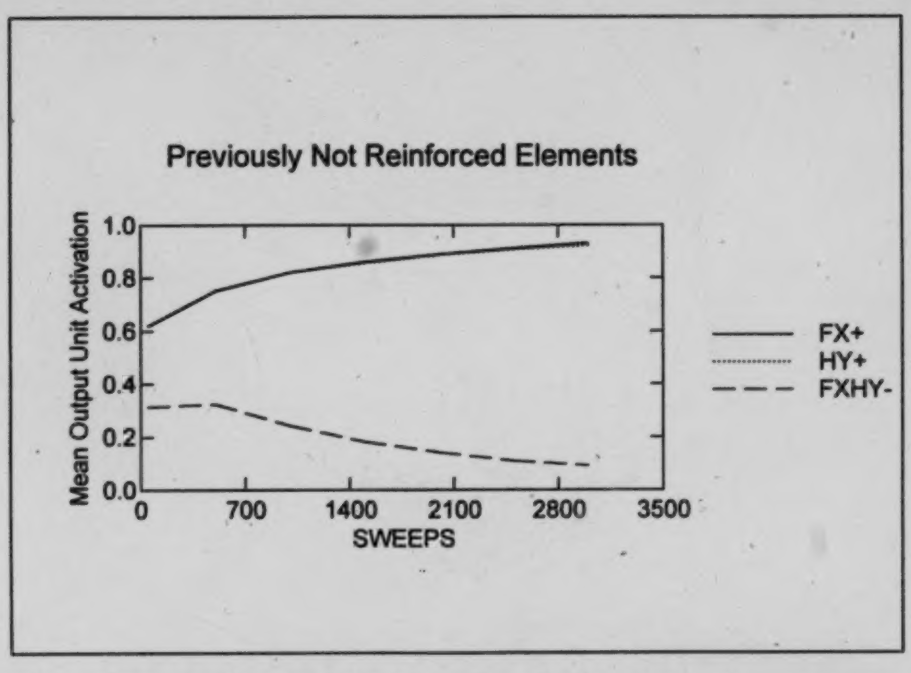

Figure 4-2. Results for Simulation 1, negative patterning task, with no pretraining of the elements. Notice that the training proceeds in a straightforward manner and that no excitatory summation is present during training.

Excitatory summation was found in the previously reinforced condition for negative patterning (see Figure 4-3).

As mentioned in the earlier introduction of their model, Delamater *et al.* overcome this deficiency in the connectionist model by using a model that has direct network connections between the input and output layers that bypass the hidden layer, in addition to the connections between the units in a layer and each unit of adjacent layers. This model produces initial excitatory summation in the negative patterning task when it is not exposed to a pretraining manipulation. This type of a move significantly increases the computational power of a network. While increases in power are often desirable, a

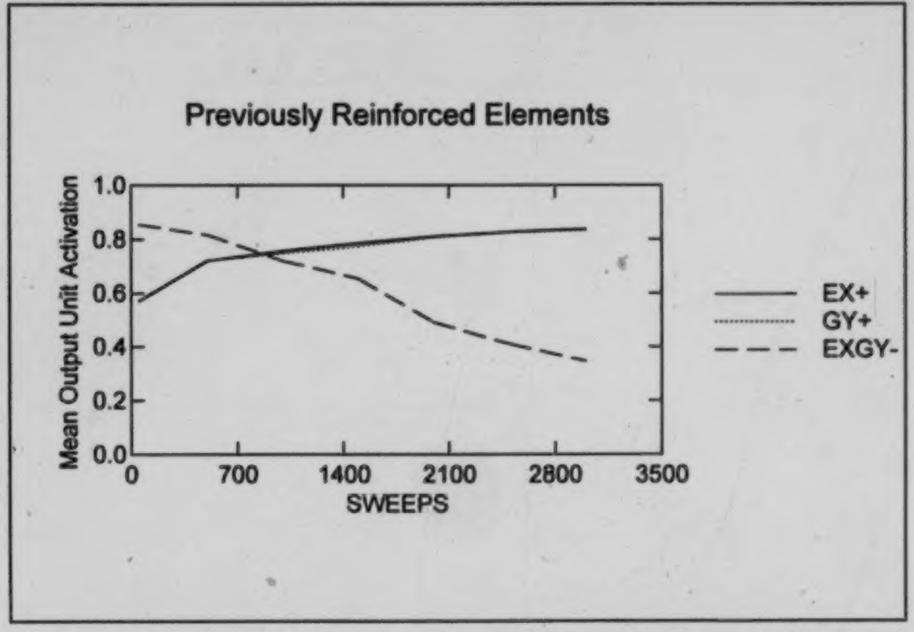

Figure 4-3. Results for Study 1, negative patterning task, with pretraining of the elements. Notice that there is excitatory summation in this condition.

problem as simple as negative or positive patterning may underwhelm a huge network. This model, although it produces the desired result, is likely unnecessarily complicated for the task it is asked to perform.

An alternative to creating a network that overfits the patterning task is to attempt to fit the data by adjusting one of the other degrees of freedom in a simulation. As discussed in Chapter 3, connectionist networks are made of many components. Delamater *et al.* altered the connectivity of the network to fit the data. In Study 1 of this thesis, the number of processing units in the hidden layer was adjusted. In the following study, it is demonstrated that changing the activation function of even a single processing unit can have an effect on the behaviour produced by that network.

Study 2 demonstrates that a relatively complex task – simultaneous negative and positive patterning training – can be performed by a simple network, as in the previous study with the simple modification of an activation function, and can produce initial excitatory summation when there is no pretraining manipulation.

Study 2: Initial Excitatory Summation

Method

Networks

The model is a connectionist network containing primarily integration device processing units (processing units with logistic activation functions) and one value unit. The processing units are arranged in three layers: an input layer of six binary units in which the input patterns are represented and a hidden layer of three integration devices. The output unit is a value unit, that is, it has an activation function that is non-monotonic. This value unit has a gaussian activation function (Dawson & Schopflocher, 1992; Dawson, 1998) that has the effect of transforming the incoming signal such that intermediate signals produce the highest output for the network, and extreme (either high or low) incoming signals produce lower activation in the output unit. The network is depicted in Figure 4-4.

Input Coding

The training set was the same as that used to train the networks in Study 1. It combined the negative and positive patterning tasks. The difference is that, given that there was no pretraining manipulation in this task, four of the elements were unnecessary. Elements A and B were always paired with background stimulus X. Elements C and D

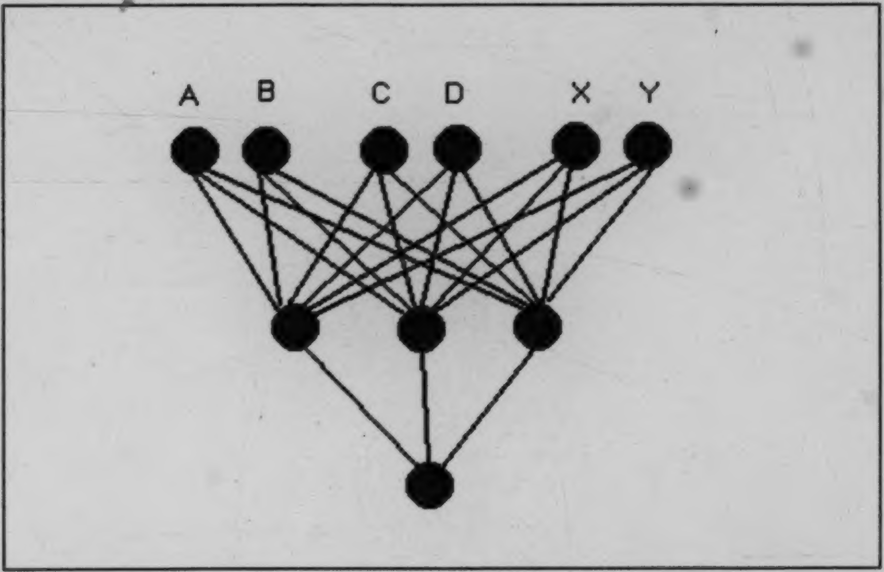

Figure 4-4. Network used in Study 2. A and B are elements used in negative patterning training; C and D are elements used in positive patterning training; X and Y are background stimuli (see text for details).

were always presented with background stimulus Y. Positive patterning contained input patterns AX^0, CY^0, $AXCY^+$; negative patterning contained input patterns BX^+, DY^+, $BXDY^0$.

Training

The network was trained with a variation of the generalized delta rule (Dawson & Schopflocher, 1992) that is designed to train networks containing value units. The learning rate was set at .03 with no momentum. The criterion for a "hit" was .1. Connection weights prior to training were randomly assigned, between the values of -.5 and .5. Weights and biases were updated after each pattern presentation, and order of presentation of the patterns was randomized.

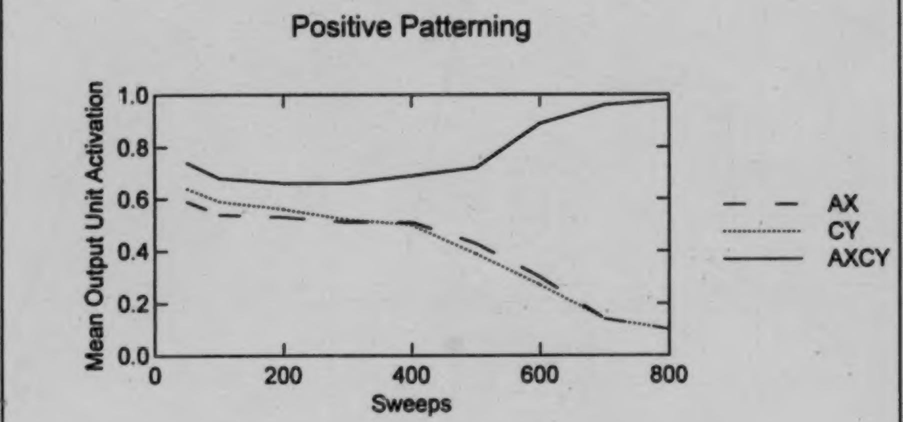

Figure 4-5. The course of learning for the positive patterning schedule in Study 2.

Results

After 838 epochs, the network converged. Learning was measured during training by recording the activation of the output unit every 50 epochs. The course of learning is shown in Figure 4-5 for positive patterning, and Figure 4-6 for negative patterning.

In the positive patterning situation with this network, I observe initial high responding by the network for both stimulus elements and the stimulus compound. The network reduces responding to all stimuli before making a strong discrimination. The network always responds, appropriately, more to the reinforced compound than to the unreinforced elements.

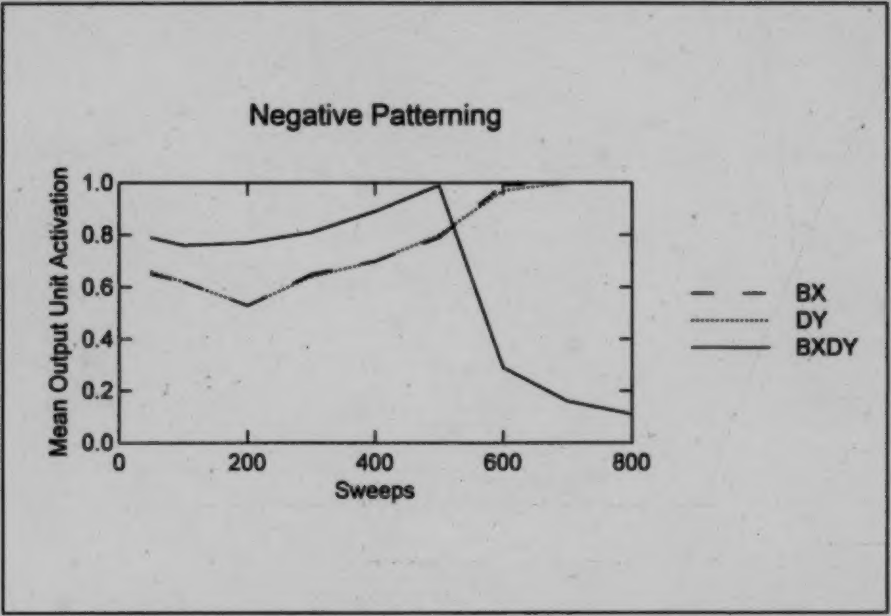

Figure 4-6. The course of learning during the learning of the negative patterning schedule.

In the negative patterning situation, the network also begins with high responding to the stimulus elements and the compound. Until around 550 sweeps, however, the network responds more to the unreinforced compound than to either of the reinforced elements. In other words, the network displays initial excitatory summation.

Discussion

It was mentioned earlier that internal representations are "semantics-free" – that is, they have no "handwired" theory in the hidden layer. The input-output mapping into which a network settles can sometimes be semantically interpreted however, by the extraction of rules from the hidden layer (Berkeley, *et al.*, 1995; Christiansen & Chater, 1992; Hanson & Burr, 1990; Dawson, 1998; Dawson, *et al.*, 1997; 2000; Elman, 1990;

Willson *et al.*, 1999). These interpretations give information about how a network is solving or has solved a particular problem and can be useful when looking for processes at work in a system. To answer process-type questions like the ones above, we need to look at the pattern of activation across the hidden units to determine what the network is doing and to see whether a semantic interpretation can be made.

Interpreting the network

Selected hidden unit activations for the negative patterning part of the problem are presented in Table 4-1. No hidden unit is able to key in on a particular pattern by the end of training, although hidden unit 3 is *primarily* responsible for representing the compound stimulus, BXDY. Given this finding, it is tempting to suggest that hidden unit 3 has emerged as a configural unit and that the representations of the elements have been distributed across the other 2 units in the network. After the network has converged, the internal activation of hidden unit 3 is .638 when BX+ is presented, .705 when DY+ is presented, and .788 when the compound BXDY- is presented. This finding could be viewed as consistent with the configural theory of Pearce (1987, 1994) in which presentation of the elements triggers a response in "configural" hidden unit 3 because of their similarity to the compound (Gluck, 1991).

If the representations were contained only within the hidden units, however, we could expect to see some marked changes at approximately 500 sweeps in hidden unit activations, given the marked behaviour change of the network at this stage of training. Instead we see relatively consistent (and small) changes in the internal activity of the units between 500 and 600 sweeps. This finding illustrates the fact that representations in PDP

Sweeps	Hidden Unit #1 BX+	Hidden Unit #2 BX+	Hidden Unit #3 BX+
100	.296	.366	.477
500	.156	.266	.543
550	.174	.260	.567
600	.195	.264	.601
838	.216	.275	.638

Sweeps	Hidden Unit #1 DY+	Hidden Unit #2 DY+	Hidden Unit #3 DY+
100	.311	.314	.564
500	.183	.212	.632
550	.206	.197	.654
600	.238	.194	.682
838	.292	.198	.705

Sweeps	Hidden Unit #1 BXDY-	Hidden Unit #2 BXDY-	Hidden Unit #3 BXDY-
100	.159	.211	.532
500	.040	.099	.636
550	.042	.082	.677
600	.045	.073	.732
838	.048	.066	.788

Table 4-1. **Selected hidden unit activations for negative patterning in Study 2.**

models are contained not just in the processing units in the network, but also in the connections between the units. In this particular case, the representations are not simply elemental or simply configural, although evidence for behaviour driven by both of these processes is in evidence in this network: the representations are distributed.

General Discussion

The studies reported in this chapter have demonstrated that psychologically interesting results can be drawn from networks with fewer hidden units than input patterns to discriminate. These studies are meant to demonstrate that it may be possible for a model of discrimination learning to conform to the packing constraint by using distributed representations – that is what PDP models are primarily designed for. But what does this mean for theories of discrimination learning?

Given that the network used in Study 2 fits the experimental data reasonably well, compared to a model that requires the addition of direct connections between the input layer and the output layer and local representations within the hidden layer, I suggest that, within this limited domain, these extra connections and extra processing units may be unnecessary in a model of discrimination learning. Without any "elemental" connections, the network looks much more like the configural model of Pearce. It is, however, a configural account with a difference since there is no longer a one-to-one mapping of stimulus pattern to hidden unit.

As mentioned previously, Kehoe and colleagues (Bellingham, *et al.*, 1985) have noted the advantages of elemental models that can explain summative processes in networks and stimulus generalization. The network presented in Study 2 of the present chapter clearly demonstrates summation at an appropriate stage of acquisition of the negative patterning task. Elemental processes, however, seem unable to account for the learning of problems that require differential associations of the elements and their compounds as is the case in positive and negative patterning. The network presented in Study 2 is able to represent elements separately from their compounds, given that it is able to solve patterning problems.

Given the present data and other sources (Kehoe, 1988; 1998; Schmajuk, 1998; Schmajuk & DiCarlo, 1992; Delamater *et al.*, 1999), it seems likely that both elemental and configural processes contribute to the learning of complex discriminations. These processes may be important at different stages of acquisition in a connectionist network: elemental processes may be important during the initial phase of training and give way to configural processes later in training. It is well known that connectionist models are dependent on the process of summation. The equivalence of the Widrow-Hoff learning rule (Widrow & Hoff, 1960/1988) used to train two-layer networks and the summation-based Rescorla-Wagner model (Rescorla & Wagner, 1972) bears this out (Kehoe, 1998; Sutton & Barto, 1981). We know, however, that the process of acquiring a behavioural response to a presented stimulus requires at least two phases: an encoding of inputs stage and a conversion of the coding to a behavioural outcome. This requires a multi-layer solution (see Chapter 3). While inputs to a particular node in a network are generally added together, that sum is then transformed according to an activation function that may be a step-type threshold function or may be a non-linear function as is used in the output unit of network presented in Study 2 of this chapter. So, although PDP models are based upon summative principles at the level of micro-processing in a particular node, the input-output mapping contained in a network generally bears little resemblance to simple summation.

While elemental processes are in evidence, the connectionist models presented in this chapter clearly represent compounds as distinct from their components – that is, they contain configural representations. They are different from other configural models, though, in the nature of the representations that are distributed across the layer of hidden units, rather than being locally coded in a single unit.

The intent of this chapter is to encourage exploration into the notion of distributed representations in models of learning. Networks that contain distributed rather than purely local representations can provide models of learning with representational power, without violation of the packing constraint. These networks have many advantages aside from the power that they conserve that have not been discussed in detail in the present chapter, however. They are less brittle when damaged and are sensitive and adaptable to changes in the environment (Hinton, *et al.*, 1986). Networks that contain distributed representations also show stimulus generalization, which is a phenomenon that initially motivated Pearce's interest in configuration (1987). Stimulus generalization happens in distributed networks because of the fact that (by definition) similar representations overlap in the network (Hinton, *et al.*, 1986; Rumelhart & Todd, 1993): when the compound stimulus AB is presented, it partially stimulates the representations of A and B that are coded across some of the same hidden units.

Networks with distributed representations may have a serious disadvantage, however. Networks that are highly distributed may be particularly susceptible to a phenomenon that has been called *catastrophic forgetting* or catastrophic interference (Ratcliff, 1990; Robins, 1995). The problem of catastrophic forgetting in networks will be explored in the following chapter.

Chapter 5

CATASTROPHIC FORGETTING

An issue in connectionism has been a tension that exists between stability and plasticity (Grossberg, 1987; Robins, 1995) in networks. A network is a model of learning, as we have seen, and of memory – in terms of the representations that are produced within the network. In this sense, it is important that these representations be preserved over time – that they be relatively *stable*. However, as models of learning, they must also be flexible or *plastic* enough to deal with new inputs and situations. These two processes are equally important in a model of learning and representation but are at odds with one another: as a network's stability increases, its plasticity decreases and vice versa (Robins, 1995, pp. 123-124).

Perhaps because of the strong relationship between connectionist models and theories of learning, as outlined in Chapter 3, stability often suffers in favour of highly plastic learning machines. Given this bias, highly plastic networks are often prone to a phenomenon called *catastrophic forgetting* or *catastrophic interference*[5] (Carpenter, 1997; Carpenter & Grossberg, 1988; French, 1992; 1997; Lewandowsky, 1991; Lewandowsky & Li, 1995; McCloskey & Cohen, 1989; Ratcliff, 1990; Robins, 1995; 1996) that arises from their lack of stability. Catastrophic forgetting occurs when the learning of new information by a network causes old information to be lost. This has been a serious setback for connectionism.

---

[5] I will be using the term *catastrophic forgetting* for the sake of simplicity throughout this paper, rather than *catastrophic interference*.

While catastrophic forgetting is a setback for a simulation of the behaviour of organisms, representational failures within a system "are frequently more important than successes, because failures can reveal critical constraints on the modeled natural system" (Pavel, 1990). This is, indeed, part of the process of learning from learning machines – interpreting the relative successes and failures of a simulation may lead us to understand more about the natural system that it models. However, any model of learning and memory that cannot overcome this simple issue of representation must be deemed a questionable model, especially given the argument that connectionist models should be considered more than an implementational level account of cognition.

Particularly prone to this problem are networks in which there are distributed representations (French, 1992; 1997), that I argued in favour of adopting as models of learning in the previous chapter. How can this dilemma be solved? Distributed representations offer much to connectionist models of learning – does there need to be a tradeoff between interesting and powerful networks with distributed representations and networks that can handle stable representations while continuing to learn in new situations? This question is explored in the rest of this chapter as I consider, first, why there is loss of information in networks, how the question of catastrophic interference has traditionally been studied and whether the forgetting is reasonable or catastrophic in discrimination learning paradigms. I will also consider a set of simulations in which I demonstrate remarkable *savings* in networks, rather than catastrophic forgetting. Studies 3-A and 3-B look into the question of the measurement of forgetting. How is forgetting best measured for discrimination tasks? Study 4 is an investigation of what is required of the problem to produce forgetting in networks. Studies 5-A and 5-B are an attempt to replicate studies that indicate that distributed networks are more prone to catastrophic forgetting than are other, more local networks.

Why do Networks Forget?

First, let us consider the question of why information is lost in connectionist networks. If we think of a neural network as a function approximator that maps a set of inputs to a set of outputs according to a particular function, $F(i)_1$. When new information is encountered, $F(i)_1$ must be altered to account for the new information, creating $F(i)_2$. If the new function, $F(i)_2$ is similar to $F(i)_1$, it is unlikely that data coded by the first function will be mis-classified according to the new function, $F(i)_2$. If the two functions are very different, however, it is likely that there will be significant interference (see Figure 5-1; from Robins, 1995, p. 136). One would expect, then, that if the intervening task is quite similar to the task in situation *A*, little catastrophic forgetting should be seen. Conversely, if the intervening schedule is very different from the schedule in situation *A*, or is opposite to the schedule in situation *A*, we should see much catastrophic forgetting.

For the studies in this section, I have selected two discrimination learning tasks that I have described in greater detail in other chapters: negative and positive patterning. These patterning schedules are particularly appropriate for a study of catastrophic forgetting, because the two tasks are almost opposites of each other. We should see high forgetting of situation *A* (negative patterning) after networks are trained on situation *B* (positive patterning) because the function associated with situation *A* is so different than

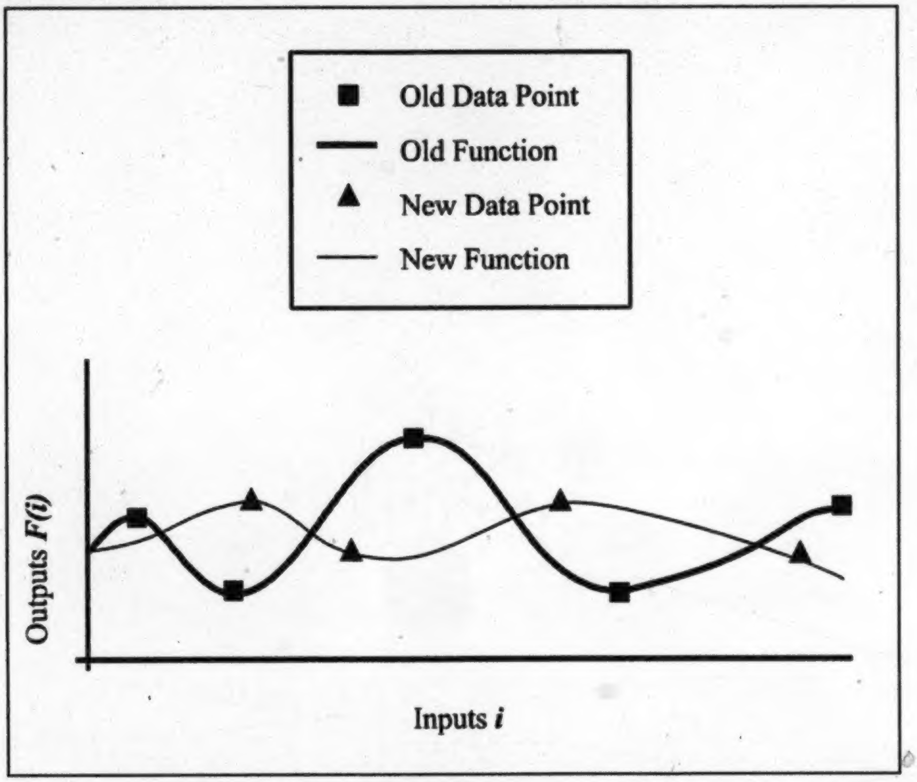
■ Old Data Point
Old Function
▲ New Data Point
New Function
Outputs F(i)
Inputs i

Figure 5-1. Neural networks as function approximators: When functions learned in Situation "a" are very different than functions learned in Situation "b", loss of fit with data points in Situation "a" is likely.

the one associated with situation *B*.

Study 3-A

Forgetting or Savings: Rate of Re-acquisition

This study, and the ones that follow, are part of an exploration into forgetting in networks. Studies 3 and 4 explore the notion that the term *catastrophic* forgetting may be a little dramatic for the forgetting phenomenon that we see in networks, at least where discrimination learning tasks are modeled. We consider, in all studies in this chapter, a set of networks trained first on negative patterning, then on an interfering task, positive patterning, then tested for recall of negative patterning. The first group of networks, for this study, is compared to a control group of networks, trained only on the last two phases: positive patterning, followed by negative patterning. This study begins by asking how forgetting might be measured appropriately when these learning tasks are considered.

How is Forgetting Measured?

The cognitive process of forgetting has aroused much attention in cognition and learning and has, no doubt, been measured many ways. In terms of the catastrophic forgetting literature, however, some conventions have emerged. Ratcliff (1990) and Robins (1995), for example, describe forgetting in terms of a loss of "goodness" of a base population of instances as new instances are introduced. They train a network on a population of base items. They then test for recall of these original items as new items are introduced, one per intervening trial. Ratcliff (1990) notes variables that influence forgetting functions, such as the size of the base population. Both Ratcliff (1990) and Robins (1995) suggest a type of rehearsal mechanism to maintain the goodness of the base population. Other solutions to the stability problem in networks have been proposed (e.g. Ans & Rousset, 2000; Grossberg, 1987; McClelland, 2000; McClelland, McNaughton, & O'Reilly, 1995). Most involve a kind of short term memory function like rehearsal or network switching that further complicates already relatively complex models of memory.

Testing forgetting in a network by looking for a loss of goodness seems specially suited to a situation in which a researcher is interested in a task in which items are added to an existing lexicon, or added to an existing list of paired associates (McCloskey & Cohen, 1989). It may be less suited to test the type of memory for schedules that learning researchers expect to find in discrimination learning.

Is This Forgetting Reasonable or Catastrophic?

For example, consider a preparation in which a pigeon is trained in a negative patterning paradigm to peck a key for a reward when a tone CS is present or when a light CS is present, but not when the two CS's occur together. This is situation *A*. The pigeon, then, is trained on a positive patterning problem, using the same tone and the same light. This time, the pigeon is given a reward when it pecks at the presentation of the compound tone and light CS, but not when it pecks at either the tone or the light alone. This is situation *B*. To solve the task in situation *B*, it is necessary for the tone+ and the light+ responses to be extinguished. In this situation, the first trial of situation *B* acts as the first extinction trial for situation *A*. Providing we test recall of situation *A* after training on situation *B*, we should see "forgetting" of the associations learned in situation *A*. One might expect, however, that the pigeon's reacquisition of the negative patterning task in the "test" phase (situation *C*) of the study would be quicker. Whether this is the case or not is, of course, an empirical question. Research supports this hypothesis (Napier, Macrae, & Kehoe, 1992): organisms reacquire extinguished responses more quickly than they acquire them initially.

We may expect that, while the specific representations *tone CS+*, *light CS+* and *compound CS-* are not actively available at the onset of testing, some kind of representation trace may be preserved throughout the training in situation *B*. If this is true, the pigeon should re-learn the schedule in situation *C* more quickly than it learned the schedule the first time it was exposed to the training conditions.

Forgetting or Remembering?

Perhaps, when looking at investigating learning tasks such as the discrimination learning tasks of interest in this thesis, rather than measuring forgetting by looking at amount of error produced by a network when a member of the base population is presented, we ought to measure *remembering* – and look for *savings* in reacquisition of the base population. This type of investigation of forgetting in networks is not unknown (see e.g. French, 1992) in which reacquisition is considered important. Heatherington & Seidenberg (1989) have suggested that the catastrophic forgetting associated with initial error after an interference task is shallow forgetting. They demonstrate that the networks have not forgotten the task by showing rapid reacquisition of the original task.

This method of testing memory is certainly well known in psychology in general, made popular as the savings method by Ebbinghaus (1885). Ebbinghaus would train himself on a list of syllables until he knew them to a particular degree of accuracy. He would test himself on these lists after a variable interval. He found that some forgetting had always taken place as his recall was not perfectly accurate. As a measure of how much had been retained, Ebbinghaus then measured how long it would take him to relearn the list – or how much had been *saved* through the interval. The present study considers forgetting in terms of reacquisition of the original task, by examining savings rather than error at the onset of retraining.

Method

Networks

The networks used in this study had two input units, two hidden units, and a single output unit. The two input units were directly connected to the two hidden units that were both connected to the output unit (see Figure 5-2). The connection weights were randomly started with a value between -.5 and +.5. The input units were binary units; the hidden units and output unit had logistic activation functions. Networks were excluded from analysis if they did not converge in the training phase. Networks were run until there were 25 networks in each condition.

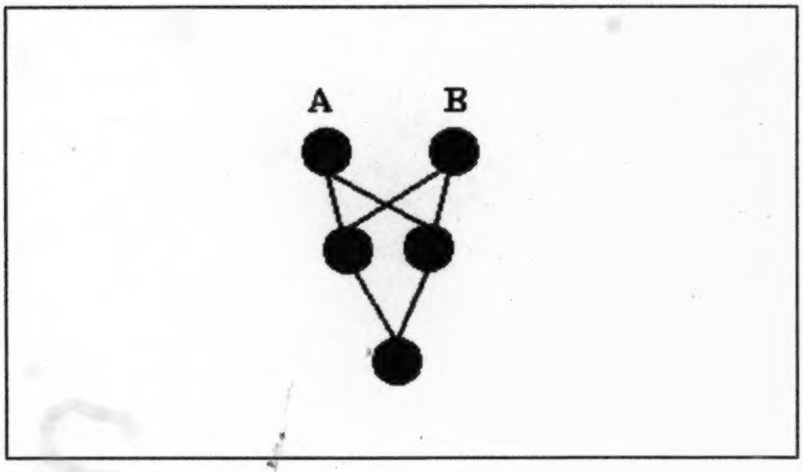

Figure 5-2. Network used in Study 3.

Training

Training consisted of three phases in the experimental condition. Phase 1 was the initial negative patterning phase. Networks were trained to respond with a 1 at the output unit, when either input element was on (1,0 or 0,1) but not when both elements were either off (0,0) or on (1,1). Phase 2 was the interference task. Networks were trained on a positive patterning schedule in which they learned to respond with a 1 at the output unit to the compound stimulus (1,1) but not to either of the two elements (1,0 or 0,1), nor when both elements were off (0,0). Phase 3 was the same as Phase 1. This condition will be referred to as the 3-phase condition.

In the control condition, networks were only trained on Phase 2 and Phase 3. This was to determine whether any facilitation seen in reacquisition of the negative patterning task in Phase 3 of the experimental condition is due to the positive patterning schedule in Phase 2 of the study, which is meant only to be an interference task. This condition will be referred to as the 2-phase condition.

All networks in this study were trained with error back-propagation (Rumelhart, *et al.*, 1986). The input patterns were presented and a difference between the network response (the activation of the output unit) was compared to an expected response for the network (either 0 or 1). The difference between response and expected response was then propagated back through the network. Training continued until the response of the network fell within .1 of the expected output for all patterns in the set, or until 20,000 sweeps. If the network had not converged by 20,000 sweeps, training was stopped and that network was deemed to have failed to solve the problem. Presentation of the input patterns was randomized within phases. Connection weights were altered after each pattern was presented. The learning rate was set at .15 for all networks. Momentum was not used.

Results and Discussion

Means and variability for number of epochs to convergence for the 3-phase condition and the control, 2-phase condition in Table 5-1. To have 25 networks in the 3-phase condition, it was necessary to run 50 networks. Of the 25 networks that were

			Phase 1	Phase 2	Phase 3	Ratio of Phase 1 to Phase 3
Study 3-A	3-Phase	mean	7288.2	1440.2	1200.5	6.2
		s.d.	2208.8	90.0	218.2	2.2
	2-Phase	mean		2648.0	5695.6	
		s.d.		218.5	1960.2	
Study 4 Discrete Elements		mean	9318.4	894.6	278.1	37.5
		s.d.	2659.1	363.9	81.2	17.5
Study 5-A 4 Hidden Units		mean	7017.1	6165.9	1227.2	6.0
		s.d.	1365.6	3650.1	281.4	1.5

Table 5-1. Means and Standard Deviations for Studies 3-A, 4 and 5-A.

excluded from the condition, 12 of them did not converge on the initial negative patterning problem and 13 failed to solve the intervening, positive patterning task.

In comparing Phase 1 and Phase 3 in the 3-phase condition, it is evident that the homogeneity of variance constraint for parametric analysis was not met [$F(24,24) = 102.7; p < .01$]. The Kolmogorov-Smirnov non-parametric test was used to evaluate the difference between these means.

In the 3-phase condition, a significant difference was found between initial acquisition of the negative patterning task in Phase 1 and the re-acquisition of that task in Phase 3 ($D_k = 1.00; p < .01$). The networks learned the negative patterning task faster the second time.

To demonstrate that this effect is not due to facilitation by the positive patterning task in Phase 2, the control condition was run (the 2-phase condition), in which networks were only exposed to Phase 2 and Phase 3. To acquire 25 networks in the two-phase condition, 88 networks had to be trained. Of the 63 networks that were excluded from this analysis, all 63 solved the positive patterning task but failed to converge on the subsequent negative patterning task. One network in the 2-phase condition was excluded from the analysis as it was identified as an outlier (epochs in Phase 3 = 19, 446; see box plot in Figure 5-3). Note that, while in the first condition, 50 networks were needed to acquire 25 networks that solved the problem (50% convergence), in this condition, 88 networks were needed to acquire 25 converged networks (28% convergence). Rather than facilitating negative patterning learning, positive patterning seems to block the learning of negative patterning when there is no prior learning, such as occurs in Phase 1 of this

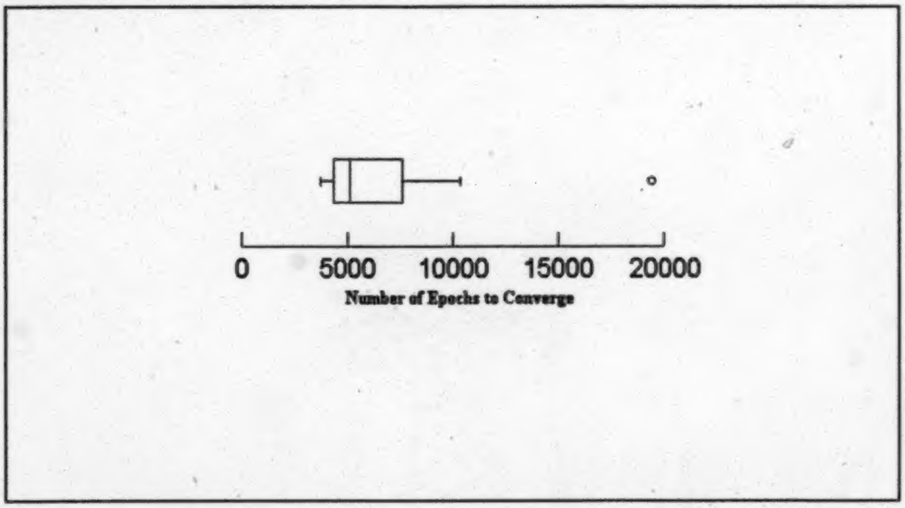

Figure 5-3. Box plot identifying an outlier in the 2-Phase condition of Study 3-A.

study, on a negative patterning task. This finding is consistent with experimental data. Bellingham and colleagues (1985) demonstrate that positive patterning blocks the learning of negative patterning in appetitive learning in rats and in aversive learning (using the nictitating membrane response) in rabbits. They found that, although positive patterning blocks negative patterning, the reverse is not true: negative patterning does not block positive patterning.

There is a significant difference between the 3-phase group's reacquisition of the negative patterning task in phase 3, and the 2-phase group's initial acquisition of the negative patterning task in phase 3 ($D_k = 1.0$; $p < .01$). Again, these groups fail to meet the homogeneity of variance constraint [$F(23,24) = 80.7$; $p < .01$]; a Kolmogorov-Smirnov analysis was used for assessing this difference. This indicates that the faster acquisition of negative patterning in Phase 3 by the 3-phase group is not due to facilitation from the positive patterning schedule. In fact, the positive patterning task seems to inhibit negative patterning learning. It must then be due to something that has been *saved* from the first learning of negative patterning in Phase 1.

Study 3-B

Forgetting or Savings: What Does Remembering Look Like?

Statistical significance aside, what course does the reacquisition of the task take? Is there enough loss between Phase 1 and Phase 3 to suggest catastrophic forgetting of Phase 1? It is useful to look at the course of learning, in this situation.

Method

Networks were the same as in Study 3-A. They had two input units, two hidden units, and a single output unit. There were initially 10 networks in each condition of this study. Networks that did not converge by 20, 000 sweeps were excluded from the analysis; this is discussed in the results section below.

As in Study 3-A, in the experimental group, networks were trained on negative patterning in the initial phase of training. In the second phase of training, networks were trained on positive patterning as an interfering task. In the final phase, networks were re-taught negative patterning. This is referred to as the 3-phase group. As in Study 3-A, in the control group, networks were trained on positive patterning, then on negative patterning. This group is referred to as the 2-phase group.

The networks in this study were trained with back-propagation. The learning rate was set at .15 for all networks. Momentum was not used.

The networks were monitored during training. The sum of squared error (SSE), summed across all four input patterns, for each network was recorded every 10 epochs.

Results and Discussion

In the experimental condition, of the ten subjects initially run, Subjects 7 and 10 were rejected from the analysis for failing to converge in Phase 1. Subjects 5 and 8 were rejected for failing to converge in Phase 2. This leaves six networks in the 3-phase

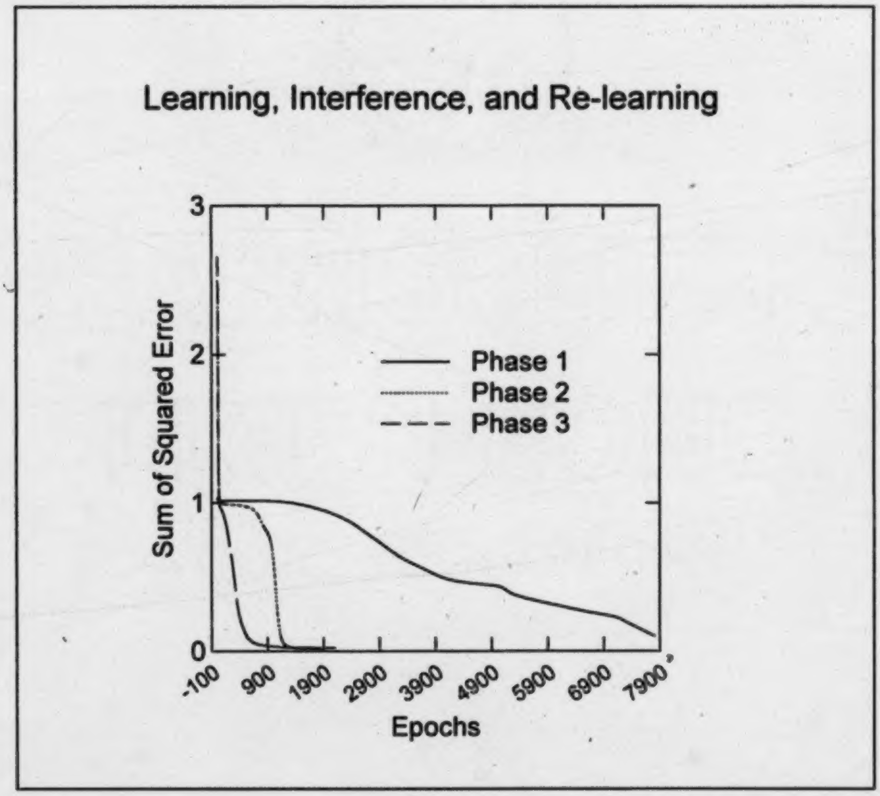

Figure 5-4. SSE plotted against number of epochs for the 6 networks in the two inputs, two hidden units, 3-phase condition, Study 3-B.

condition in this study. In the control condition, of the ten networks initially run, all ten converged on the positive patterning task, but only four were able to go on and solve the negative patterning task. The six networks that failed to solve negative patterning in the 2-phase condition were excluded from analysis.[6]

Figure 5-4 shows the average SSE plotted against the number of epochs for the six networks in the 3-phase condition, for each of the 3 phases.

Notice the initial average SSE in Phase 3 of the 3-phase condition. If we were measuring forgetting as Ratcliff (1990) did, we would likely conclude that the networks had forgotten the task. The course of learning demonstrates, however, that the task has not been forgotten. Information about the negative patterning schedule has allowed the network to solve the task more quickly in Phase 3, in spite of the fact that the responses acquired in Phase 1 would have had to be extinguished in Phase 2.

Figure 5-5 shows the average SSE for the 4 networks in the 2-phase condition, for both phases. The phases in this graph are labeled *Phase 2* and *Phase 3*, even though the networks in this condition never received a *Phase 1*, for ease of comparison. Phase 2 is acquired more quickly than Phase 3 – positive patterning is a simpler problem than negative patterning. The significant effect to note here, however, is the difference between Phase 3 in the 3-phase condition (in Figure 5-4) and Phase 3 in the 2-phase condition.

---

[6] While the *n*'s for this study seems small (6 and 4), their size is justified given that the point to be made here is a simple one, and does not require a means comparison. Additionally, the wiretapping procedure of watching a network train generates a vast amount of data and is resource intensive.

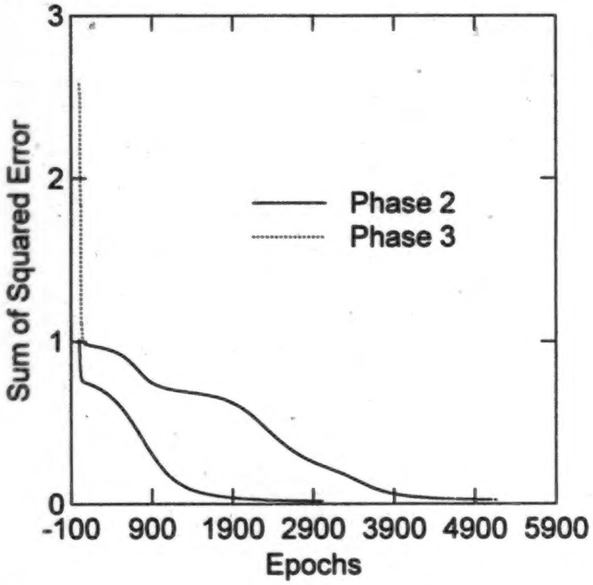

Figure 5-5. SSE for the 4 networks in the two inputs, two hidden units, 2-phase condition, Study 3-B.

In Study 3-A, it was demonstrated that a statistically significant difference exists between the acquisition of negative patterning in Phase 1 and reacquisition of negative patterning in Phase 3, and that this effect was not due to priming from the positive patterning, interference task. In this study, it is demonstrated that this statistical significance translates into practical significance or *apparent significance* as shown in the figures: networks re-learn negative patterning after an interference task faster than they learn it initially.

When forgetting is measured in terms of savings, the outcome seems reasonable, not catastrophic. As mentioned earlier, an organism would certainly require a number of trials to determine what type of schedule was being presented, particularly if the elements were the same across phases of the experiment. What if the elements were not the same? An experimenter would expect, in this case, that the organism would require less time to reacquire an initial task if the interfering task did not require the extinction of a particular response from that task. Study 4 explores the notion of savings using separate elements for the intervening task.

Study 4: Separate Elements, Separate Tasks

Method

Networks

Networks were run until there were 25 networks in each condition. All networks had four input units. Two of these input units, A and B, were used for Phase 1 and Phase 3 (negative patterning) while the other two, C and D, were used for Phase 2 (positive patterning). The networks in this condition had two hidden units and a single output unit (see Figure 5-6) All processing units had logistic activation functions. Before training the connection weights were randomly started between -.5 and +.5.

Input Coding

The training set reflects the intention to use separate inputs for negative and positive patterning. The input patterns were coded across the 4 input units that had activations of 0 or 1. Phase 1 consisted of a negative patterning problem in which the elements A (1,0,0,0) and B (0,1,0,0) were reinforced, while the AB compound stimulus

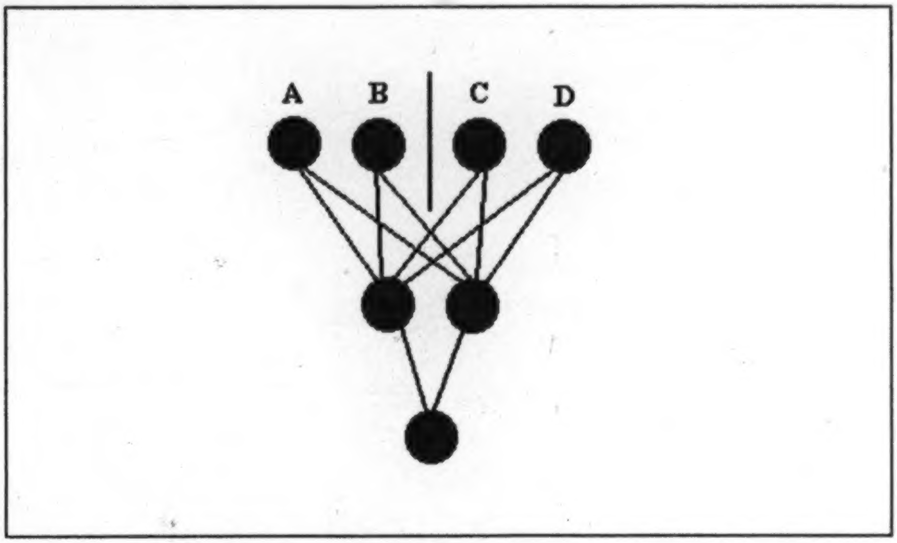

Figure 5-6. Network used in Study 4. This network is the same as that used in Study 3, except that the input patterns are coded across separate input unit pairs, depending upon the phase of training (see text for details).

(1,1,0,0) was not. Phase 2 consists, as in the previous study, of a positive patterning schedule in which elements C (0,0,1,0) and D (0,0,0,1) were not reinforced while the CD (0,0,1,1) compound was reinforced. A null input pattern (0,0,0,0) was presented without reinforcement in each of these phases. Input patterns for Phase 3 were the same as in Phase 1.

Training

All networks were trained as in Study 3-A and 3-B, using back-propagation (Rumelhart, *et al.*, 1986). Input pattern order was randomized. Connection weights were updated after each pattern presentation. The learning rate was set at .15 and the networks were trained without momentum. Networks were trained on the Phase 1 inputs until the response of the network fell within .1 of the expected output state for all patterns. The networks were then trained on Phase 2 inputs, then on Phase 3 inputs. If a network had not found a solution within 20,000 sweeps in any of the three phases, training was stopped.

Results and Discussion

Means and variances are in Table 5-1 (p. 95). To acquire a subject base of 25 networks per condition, 34 networks were run. Homogeneity of variances was violated for a comparison of means in Phase 1 and Phase 3 [F(24,24) = 1072.4; $p < .01$]. A Kolmogorov-Smirnov one-sample test was done to evaluate group differences. The difference between number of sweeps to convergence in Phase 1 and Phase 3 is significant ($D_k = 1.0$; $p < .01$): there is considerable savings from Phase 1 to Phase 3.

The task in this study is initially more difficult with four input units than the task in Study 3 in which there are only two input units (means are 7288.2 for the two-input problem and 9318.4 for the four-input problem in Phase 1), since changing the number of input units changes the dimensionality of the problem. Between these groups, variances are homogenous [F(24,24) = 1.4; $p > .01$], so a t-test was done to evaluate the difference between means for Phase 1 acquisition [t pooled (48) = 2.9; $p < .01$]: the four-input problem is more difficult than the two-input problem, as measured by the rate of acquisition of the task.

To account for this difference in difficulty, ratios of number of epochs in Phase 1 to number of epochs in Phase 3 were calculated for the two-input and the four-input networks. Average ratios of Phase 1 to Phase 3 acquisition for the two-input condition presented in Study 3 and for the four-input condition presented in this study are recorded in Table 5-1. The difference between ratios in the two-input condition and the four-input condition were evaluated with a Kolmogorov-Smirnov non-parametric test as these groups fail to meet the homogeneity of variance constraint [$F(24,24) = 63.3$; $p < .01$]. The difference between means of ratios is significant ($D_k = 1.0$; $p < .01$). This indicates that, after a correction has been made for scale differences, the Phase 3 rate of learning is closer to the Phase 1 rate of learning in the two-input condition than in the four-input condition. As predicted, there is significantly less forgetting when the elements are either used for one task or the other but not for both.

Study 5-A

Local and Distributed Representations: Savings

As mentioned previously, the research notes that particularly prone to catastrophic forgetting are networks that have distributed representations (French, 1992). This is particularly relevant to the argument made in the previous chapter that distributed representations provide more interesting and powerful accounts of cognition than local models. Clearly if they are more prone to unnatural forgetting, they are less plausible models of cognition than their local counterparts. I attempt to replicate this finding in this study.

One proposed means of manipulating the locality of the representations in a network is by altering the number of hidden units (Seidenberg & McClelland, 1989). In this study, networks from Study 3-A were presumed to contain more distributed representations and were used as a "distributed" comparison group. Given that there are four input patterns and only two hidden units in these networks, individual hidden units cannot be trained to locally key in on a particular input pattern. Representations must be distributed in these networks. In the "local" condition in this study, networks were provided with as many hidden units as input patterns. Networks were not forced into developing local representations of the inputs. Rather, local representations were *permitted* in the networks, given that the architecture allowed for them and networks tend to be lazy in that they generally choose the simplest representation that their architecture allows. This study evaluates the effect of providing the network with an opportunity to produce local representations. It is concerned with forgetting from a savings perspective – rate of reacquisition of the task is measured. The following study, Study 5-B, looks at forgetting in a manner more similar to that of Ratcliff (1990) – in terms of SSE. Study 5-C is concerned with whether reducing the number of hidden units in the network has, indeed, produced networks with more local representations than the networks in Study 3-A.

Method

Networks

The study was run until there were 25 networks in each condition. These networks were similar to those used in Study 3-A and 3-B, but they had two input units and four hidden units. Input units were connected to all four hidden units; the hidden units were connected to a single output unit. All processing units had logistic activation functions. Connections were randomly started between -.5 and .5.

Training

Training was the same as that in the 3-phase condition of Study 3. Networks were trained on a negative patterning problem in Phase 1, a positive patterning problem in Phase 2, and retrained on the negative patterning task in Phase 3. There were four input patterns per phase (0,0; 1,0; 0,1; 1,1) with expected outcomes governed by the schedule (negative or positive patterning) for each phase.

Networks were trained using back-propagation. The order of presentation of inputs within phases was randomized. Connection weights were updated after each pattern was presented. The learning rate was set at .15. Networks were trained without momentum. Training was stopped when output was within .1 of the expected output for all patterns in the training set, or when the network reached 20,000 sweeps without finding a solution to the problem.

Results and Discussion

Means and variances are presented in Table 5-1. To acquire a subject base of 25 in each condition, 28 networks were run. Of the three networks that failed to find a solution to the 3-phase task, all three failed to solve the positive patterning problem in Phase 2. Variances in Phase 1 acquisition and Phase 3 acquisition are not homogeneous [$F(24,24) = 23.6$; $p < .01$]; a Kolmogorov-Smirnov test was used to compare means. The difference between acquisition of negative patterning in Phase 1 and reacquisition in Phase 3 is significant ($D_k = 1.0$; $p < .01$); the networks relearned the task in fewer sweeps than it took them to learn it the first time.

Differences between the networks in the 4-hidden-unit condition and networks in the 2-hidden-unit condition can be assessed by the same method used in Study 4 – by using a ratio of acquisition in Phase 1 to acquisition in Phase 3 for a group of networks with only two hidden units and a group of networks with four hidden units. The ratio mean and variance are reported in Table 5-1 (p. 95) for the four-hidden-unit networks. The ratios for the networks with four hidden units were compared to the ratios for the 3-phase group in Study 3-A. Variances were considered homogeneous [F (24,24) = 2.63; *p* > .01] so a *t*-test was used to evaluate the differences between means of ratios. There is no significant difference between these groups in rate of reacquisition of the negative patterning task in Phase 3 [t pooled (13) = .216, *p* > .05] when we look for differences by this method.

Savings is only one way to judge forgetting, however. While I argued earlier that savings may be a more appropriate way to look at forgetting when using the types of discrimination learning tasks used throughout this chapter, many researchers have used other measures. When comparing the two groups of networks described in this study, what happens when we look at a measure more akin to Ratcliff's (1990) *goodness* construct?

Study 5-B

Local and Distributed Representations: "Goodness"

How can we evaluate *goodness* in these networks? Comparing the total SSE for the networks in the two conditions (two hidden units and four hidden units) before any retraining takes place in Phase 3 would be similar to goodness of a base population after an intervening task.

Method

Networks

Data for the control, two-hidden-unit condition was taken from the 10 networks originally run in Study 3-B in the 3-phase condition. These networks had two input units, two hidden units, and a single output unit.

In the experimental condition in this study, networks were the same as in the two-hidden-unit condition, except that they contained four processing units in the hidden layer. They had two input units and a single output unit. All processing units had logistic activation functions. Connections were randomly started between -.5 and +.5.

Training

Networks in both conditions were trained on the first 2 phases of the 3-phase task described originally in Study 3-A. In Phase 1, networks were trained on negative patterning, until the output of the network fell within .1 of the expected response of the network for all patterns. In Phase 2, networks were trained on positive patterning until the response fell within .1 of the expected output for all patterns. In Phase 3, a measure of goodness was recorded before retraining took place. Goodness was measured as SSE of the network when the network was presented with the first trial of a negative patterning schedule, summed across all input patterns.

These networks were trained with back-propagation (Rumelhart, *et al.*, 1987). Input patterns were presented in a random order. Connection weights were updated after each pattern presentation. If networks had not found a solution to the task in either Phase 1 or Phase 2 by 20,000 epochs, they were excluded from the study.

Results and Discussion

In the control group containing the networks with only two hidden units originally presented in Study 3-A, of the ten subjects initially run, Subjects 7 and 10 were rejected from the analysis for failing to converge in Phase 1. Subjects 5 and 8 were rejected for failing to converge in Phase 2. This leaves six networks in the two-hidden-unit condition in this study.

In the four-hidden-unit condition, of the 10 subjects run, Networks 5, 6, and 7 were excluded from analysis, as they never converged on a solution to the negative patterning task. This leaves 7 networks in the study. Figure 5-7 shows average SSE

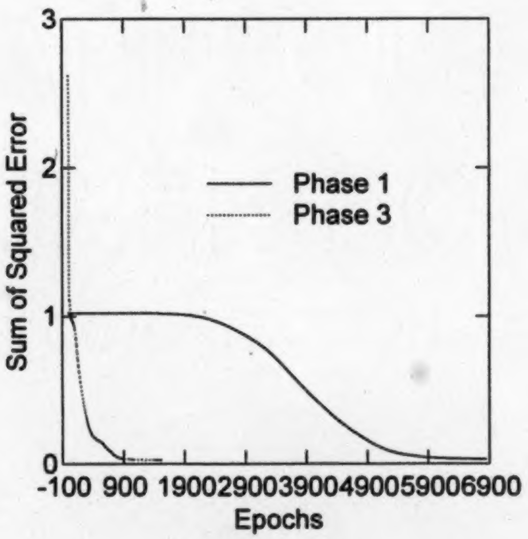

Figure 5-7. SSE for the 9 networks in the local representation condition: two input units, four hidden units, Study 5.

during training in the three phases for the 4 hidden unit networks.

Comparing SSE prior to any training on the negative patterning task in Phase 1 for the two hidden unit condition (mean = 1.025, s.d. = .008) and the four hidden unit condition (mean = 1.030, s.d. = .015) demonstrated that there was no difference between the groups initially [*t* pooled (13) = .94; $p > .05$]. In task reacquisition, however, in Phase 3, the networks in the two hidden unit condition (mean = 2.654, s.d. .019) had more loss of "goodness" than the networks in the four hidden unit condition [mean = 2.620, s.d. = .034; *t* pooled (13) = 2.17; $p < .05$].

This finding seems to provide evidence for the notion that networks that contain more local representations are less prone to forgetting a task after being trained on an interfering task than are networks with more distributed representations. This would be consistent with earlier reports of forgetting in local and distributed networks (French, 1992). Is the manipulation of local to distributed representations in these networks valid, however? Although the networks have enough hidden units to preserve each representation in a single hidden unit, to determine whether they have done this, we need to look inside the network.

Study 5-C

Local and Distributed Representations: Network Interpretation

Method

The control network in this study was the same as the networks run in Study 3-A. It had two input units, two hidden units and a single output unit. In the "local" condition, the networks had two input units, four hidden units and an output unit. There was one network analyzed per condition.

Training was the same procedure as in previous studies in this chapter and was the same for both conditions in this study. In Phase 1, the networks learned to respond appropriately to a negative patterning schedule. They were trained until their output was within .1 of the expected output for all training patterns. They were then trained on a positive patterning schedule as an interference task, again, until their output reached the criterion of being within .1 of the expected output for all patterns. The networks were then re-trained on the negative patterning schedule.

These networks were trained using back-propagation (Rumelhart *et al.*, 1986). Input patterns were presented in a random order. Connection weights were updated after each pattern presentation The learning rate was set at 1.5. No momentum term was used in this study.

Hidden units were "wiretapped" throughout training. Wiretapping is a process in which the internal activation of each hidden unit is recorded for each input pattern. Of importance to this study are the wiretaps taken during the final epoch of training in each phase. These hidden unit activations tell a story about how the representations are encoded in the hidden layer at the conclusion of training.

Results and Discussion

Hidden unit activations for the control network that contain only two hidden units are presented in Table 5-2. Distributed representations most often present as a group of non-0 weights that are not too high for each hidden unit (Hanson & Burr, 1990). However, looking, in particular, at hidden unit activations in Phase 3, we see that both

		Hidden Unit 1	Hidden Unit 2
Phase 1	Null-	.99	.88
	A+	.84	.03
	B+	.84	.03
	AB-	.13	.00
Phase 2	Null-	.82	.99
	A-	.03	.66
	B-	.03	.66
	AB+	.00	.03
Phase 3	Null-	1.00	.83
	A+	.82	.01
	B+	.82	.01
	AB-	.08	.00

Table 5-2. Hidden unit activations for the "distributed" network in Study 5-C.

Hidden Units 1 and 2 contain only relatively high activations and very low activations. Hidden Unit 1 responds strongly to the null pattern and to each of the elements, but does not respond to the compound stimulus. This unit functions within the network as a "NOT AND" detector – that is, it is tuned to not respond to the AND (1,1) pattern. The other hidden unit in this condition is an "NOT OR" detector – it responds to the compound but does not respond when either A, or B, or A and B are detected. These hidden units do not contain highly distributed representations, as expected. Rather they locally key in on particular features in the input space. These features are not as simple as a one-pattern,

		Hidden Unit 1	Hidden Unit 2	Hidden Unit 3	Hidden Unit 4
Phase 1	Null-	.89	.46	.99	.26
	AX+	.03	.72	.82	.43
	BY+	.03	.55	.83	.56
	AXBY-	.00	.79	.18	.73
Phase 2	Null-	.99	.51	.85	.36
	AX-	.67	.67	.05	.47
	BY-	.68	.52	.06	.56
	AXBY+	.03	.68	.00	.66
Phase 3	Null-	.88	.52	.99	.32
	AX+	.02	.76	.81	.55
	BY+	.02	.62	.81	.66
	AXBY-	.00	.83	.13	.83

Table 5-3. **Hidden unit activities of the four hidden units in the network in the "local" condition in Study 5-C.**

one-unit mapping that is discussed in Chapter 4, but can still be described as local representations.

Hidden unit activations for the networks that contain four hidden units are presented in Table 5-3. These activations reveal an interesting pattern, primarily in their similarity to the activations in Table 5-2. For example, Hidden unit 1 in this four-hidden-unit network is familiar. It too is a *NOT OR* detector like Hidden unit 2 in the two-hidden-unit condition. And Hidden Unit 3 in the four-hidden-unit condition is a *NOT*

AND detector – it is the same as Hidden Unit 1 in the two-hidden-unit condition. The remaining two hidden units in the four-hidden-unit condition contain distributed representations. Neither of these units "keys in" on any particular feature in the input space.

Clearly, then, the manipulation of reducing the number of hidden units in these networks has not produced networks with more distributed representations. If anything, the reverse is true. The larger network contains two hidden units that are not local feature detectors. A question to ask about these units, then, is: if they are not feature detecting, what are they doing? The answer to this seems to be that they are not responsible for much! We know that, in the four-hidden-unit network, Hidden Units 1 and 2 are sufficient to solve the problem – given that two hidden units with the same function are able to solve the problem in the two-hidden-unit condition. The other two hidden units are likely extraneous in the network.

This discussion brings us back around to where we were in Chapter 4. There is a problem in the four-hidden-unit network of *overfitting*. In Study 5-B, I presented evidence that supported the notion that there was more forgetting in the two-hidden-unit condition than in the four-hidden-unit condition, when forgetting was measured by SSE at sweep 1 in Phase 3. We now know that this was not due to a local versus distributed manipulation. I would speculate that, instead, the effect could be due to the problem of *overfitting*. It could be the case that networks that overfit the problem of interest are less prone to forgetting when goodness is measured.

General Discussion

Studies 3-A and 3B were concerned with catastrophic forgetting in standard, distributed networks, looking at savings in reacquisition instead of by error. Study 4 contained a manipulation of the architecture of a network; by increasing the number of input units, negative and positive patterning could be taught using separate elements and forgetting was significantly reduced. Studies 5A, 5B, and 5C were concerned with investigating the claim that distributed networks are more susceptible to forgetting than localist models. In these studies, network architecture was manipulated by changing the number of hidden units available to represent the problems. While an effect was found, the manipulation of local versus distributed was questioned in Study 5-C and it was suggested that the effect could be related to a problem of overfitting.

A theme across the studies in this chapter is that the phenomenon of catastrophic forgetting is dependent upon how you look at the problem, and upon what problems you look at. For certain learning tasks, it seems unreasonable to presume that networks should not have any loss when an intervening task is introduced. In this sense, the term "catastrophic" as it applies to the type of forgetting seen in networks may be an overstatement. Studies 3A and 5A, include an exploration of a different way that forgetting can be measured in networks. For the discrimination learning tasks presented in this thesis, I argue that *savings* may be more relevant than a *goodness of base population* measure with no retraining.

When savings are considered, it appears that trace representations of the original task are available to the network throughout the interference phase. This is consistent with the findings of Heatherington & Seidenberg (1989). During Phase 2 training, the negative patterning learning is suppressed so that the network can produce appropriate behaviour in the positive patterning phase. The negative patterning task is not forgotten, however. Evidence for this is provided in all conditions – reacquisition of a task is much faster than original acquisition.

There is less forgetting in networks in which the elements are unique to either the negative or the positive patterning task (Study 4). This is consistent with what would be expected: more is preserved of the original representation of the negative patterning task.

If there is less forgetting in local networks than in distributed networks (French, 1992), evidence for this is not found in this thesis. When "goodness" is evaluated, however, instead of savings, there is an effect for number of hidden units in a network: Networks with more hidden units display less forgetting than distributed networks (Study 5-B). The effect for more forgetting in these larger networks provides a weak effect, however, and is dependent upon the measure used. In addition to the marginality of the finding, SSE measure or the "goodness" measure does not seem an appropriate way to measure forgetting in this particular discrimination learning task. It is illogical to expect an organism or a network to respond appropriately to a schedule that it does not *expect*. After being trained on the positive patterning schedule in Phase 2, testing a network's recall of negative patterning before it is informed of the schedule change seems trivial. This is borne out in a look at the SSE measure for both the two and the four hidden unit conditions in Study 5-2. The SSE is significantly higher at the outset of training in Phase 3 than it was before any training took place for both conditions. This makes sense when we consider how different the functions must be to fit the inputs in the negative patterning task and in the positive patterning task. The SSE findings may be more relevant within a different research domain than they seem to be for discrimination learning phenomena.

Throughout this chapter, I have presented studies in which evidence for savings is prominent. Rather than finding *catastrophic* forgetting in networks, the studies in this chapter demonstrate that, while networks do forget, networks also *remember*, that appears more relevant to this domain of study.

In Chapter 4, distributed models were presented as necessary in models of discrimination learning. The exploration of catastrophic forgetting in this chapter came about because of a finding that, although distributed networks may be desirable from one perspective, it is possible that models that contain distributed representations are more prone to forgetting. The number of hidden units in a network is a variable – a variable about which the network designer makes a decision. The number of hidden units included in a network is not the only decision that a modeler makes when designing a network as a model of learning. The following chapter explores aspects of design in simulations that can have an influence on the behaviour produced by connectionist models.

Chapter 6

EVALUATING CONNECTIONIST MODELS OF LEARNING

Connectionist networks have been put forward as models of associative learning (Gluck & Myers, 1993; Kehoe, 1988; Shanks, 1995). Networks have been proposed as models in specific learning situations, such as conditional discrimination (Maki & Abunawass, 1991; Pearce, 1994), as has been discussed in previous chapters. Testing the plausibility of these models is certainly important if they are to be considered useful in the description of associative learning. A strategy for the evaluation of these models is to attempt to fit human or animal performance on a specific task to data from a comparison group of network "subjects".

As an example from the conditional discrimination literature, consider, again, the study by Delamater, *et al.* (1999) in which the researchers pre-conditioned rats on simple stimuli, then trained the rats using either the previously reinforced stimuli, and a compound of the same stimuli, or previously not reinforced stimuli, and a compound of those stimuli. The training situation involved either a positive patterning task (in which only the compound stimulus was reinforced) or a negative patterning task (in which either the first or second stimulus when presented alone was reinforced, but the compound was not reinforced). The purpose of the study was to evaluate the ability of three competing theories of discrimination learning to explain the performance of the animal sample. One of these theories was a connectionist theory that, most generally, was described by Delamater *et al.* as a model in which representations change during the course of learning and solutions to the task develop depending on the nature of the task.

Delamater *et al.* (1999) showed that a connectionist model demonstrated facilitation in positive patterning learning when the stimuli had been previously reinforced demonstrated facilitation in negative patterning learning when the stimuli had not been previously reinforced. However, rats that had been pre-conditioned, then trained on the task failed to follow this pattern. Instead, the rats demonstrated facilitation in both negative and positive patterning when the stimuli had been previously reinforced.

The experimental component of the Delamater *et al.* study involved rats in one of four conditions. The rats were pretrained on the same task as the networks, and were then given one of four problems to solve involving combinations of four stimuli: two distinct visual stimuli and two distinct auditory stimuli. As in the network study, the conditions were: a positive patterning problem with elements that had been previously reinforced; a positive patterning problem with elements that had been previously not reinforced; a negative patterning task with elements that had been previously reinforced; a negative patterning task with elements that had been not reinforced in the pretraining session.

The main findings in the Delamater *et al.* rat study gave the researchers reason to question the validity of the neural network model they used. The rats showed a) slight facilitation for acquiring the positive patterning task in the previously reinforced condition (consistent with the model) and also b) facilitation for acquiring the negative patterning task in the previously reinforced condition (contrary to the prediction of the model). In the negative patterning situation, the researchers also found that the rats learned the discrimination faster between the auditory component and the compound than they did between the visual component and the compound. The negative patterning situation also brought out an initial excitatory summation effect, that is, early in training, the rats responded more to the compound than to either of the elements, although it had never before been exposed to the compound, as explored in Chapter 4.

As a result of the simulation data being a poor fit to the experimental data, Delamater *et al.* concluded that "the present data suggest that if changes in the internal representations of stimuli occur throughout training, they do not do so in the manner anticipated by the standard multi-layered network model of Rumelhart *et al.*" (1999, p. 108). While their conclusion that their rat subjects do not learn like connectionist networks is certainly true of the particular network used by these researchers, it should be recognized that there are almost as many varieties of connectionist models as there are researchers who use them (Dawson, 1998, Chapter 3). The notion of testing a general connectionist theory is problematic, since there is no general connectionist network against which an appropriate hypothesis can be constructed. Hanson (1990) acknowledges that "there are no obvious principles that will allow the generic design of connectionist models at this point" (p. 512). Gluck and Bower (1988) suggest that "it is difficult to test the adaptive network framework in general. Rather, one can only test a specific realization of the framework" (p. 167). However, it is possible to develop strategies for the evaluation of information processing systems that are dependent upon the nature of the question being investigated. Gluck and Bower go on to say "by noting the circumstances where it predicts accurately versus those where it has shortcomings, we can gather generalizations about which network assumptions and learning algorithms are generally adequate to explain results across broad ranges of experimental conditions" (1988, p. 167).

A model for this evaluation process can be found in a paper by Allan Newell (1973) in which Newell attempts to model human control processes with a production system[2] using a classic cognitive psychology paradigm (Sternberg, 1970). While Newell's specific model of control is not particularly relevant to the present discussion, his scheme for fitting the data is one that is loosely followed in the rest of this chapter. Newell begins his argument with a series of questions about the production system of interest. He claims that "there are many questions that can be answered in many different ways. Each assemblage of answers yields a different production system with different properties from its siblings. Taken in all, they constitute a family of schemes for specifying information processing systems" (Newell, 1973, p. 464). Throughout the remainder of the paper, Newell presents various production systems that do not function appropriately when performing this well understood task. Production systems 1 - 6 are presented in the paper with explanations of the characteristics that they do not possess that they should. Production system 7 is presented as being "close to satisfying the several empirical propositions listed earlier" (p. 492). The successes and failures of this model led Newell to make further hypotheses about human information processing.

In the same way, any connectionist network that we could consider as a model of discrimination learning is a member of a connectionist family with a host of variable properties. On this subject, Rumelhart & McClelland (1986) comment:

---

[2] Production systems are a class of (non-connectionist) information processing systems that behave according to a set of operators or rules. The production system used in Newell's argument in the 1973 paper is explained in much detail in Newell and Simon (1972).

> "We don't really have a single model. Rather, we have a family of related models. In the best of all worlds each of our specific models may turn out to be a rough approximation to some unifying, underlying model as specialized to the problem area in question. More likely, however, each represents an exploration into a more or less uncharted region of the space of PDP models." (p. 145)

In this chapter, I explore various members of the family of connectionist models to determine their fit to the rat data in the study discussed earlier by Delamater *et al.* (1999). Network architecture is manipulated in Study 6, the structure of the problem space in Study 7, and an aspect of learning in Study 8. By manipulating these aspects of simulation, it is demonstrated that a reasonable fit to experimental data can be achieved with a neural network model.

Study 6: A Perceptron Solution to Negative Patterning

When considering models of data, it is generally considered most acceptable to begin with the simplest model that is powerful enough, in principle, to account for the data, or to solve a problem of interest. This is similar to a parsimony argument in which simpler theories are preferred over more complex ones, provided they contribute a viable solution.

In the Delamater *et al.* (1999) study, rats were trained on negative and positive patterning tasks. Previously, I have noted that negative patterning is equivalent to the logical X-OR problem, and that positive patterning is equivalent to the logical AND problem. In the simulation portion of the study, however, the researchers have used only three input patterns per condition.

The simulated learning situation of Delamater *et al.* (1999) consisted of four separate stimuli and two background stimuli. The first two stimuli (A and B) were considered distinct from one another but related in some way (as two distinct auditory cues would be). They were never presented together and, when either A or B was presented, the background cue X was presented with them. The second two stimuli (C and D) were also considered distinct yet related (as two visual cues would be). C and D shared a common background cue (Y) and were never presented together.

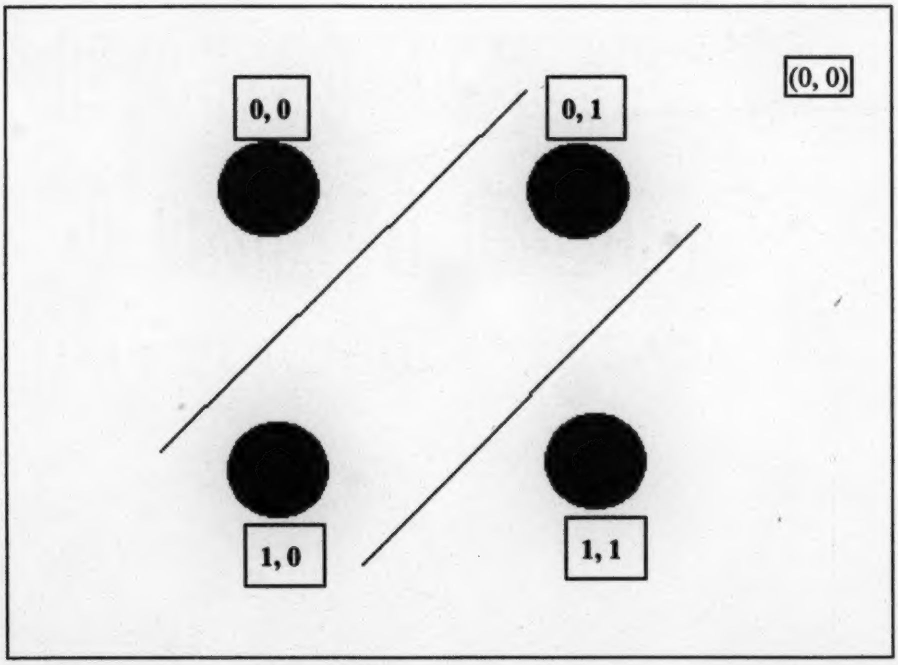

Figure 6-1. Linear non-separability, with each discrimination represented by a line.

In positive patterning the patterns presented by Delamater *et al* (1999) were AX-, CY-, AXCY+ in the previously reinforced condition and BX-, DY-, BXDY+ in the previously not reinforced condition. In negative patterning the patterns were AX+, CY+, AXCY- in the previously reinforced condition and BX+, DY+, BXDY- in the previously not reinforced condition.

In spite of the fact that the negative patterning situation (logical X-OR) is a linearly non-separable problem as discussed in Chapter 3 (see Figure 6-1), the problems in all four of these conditions are linearly separable (see Figure 6-2). While linearly non-separable problems require two discriminations or two carvings of the problem space,

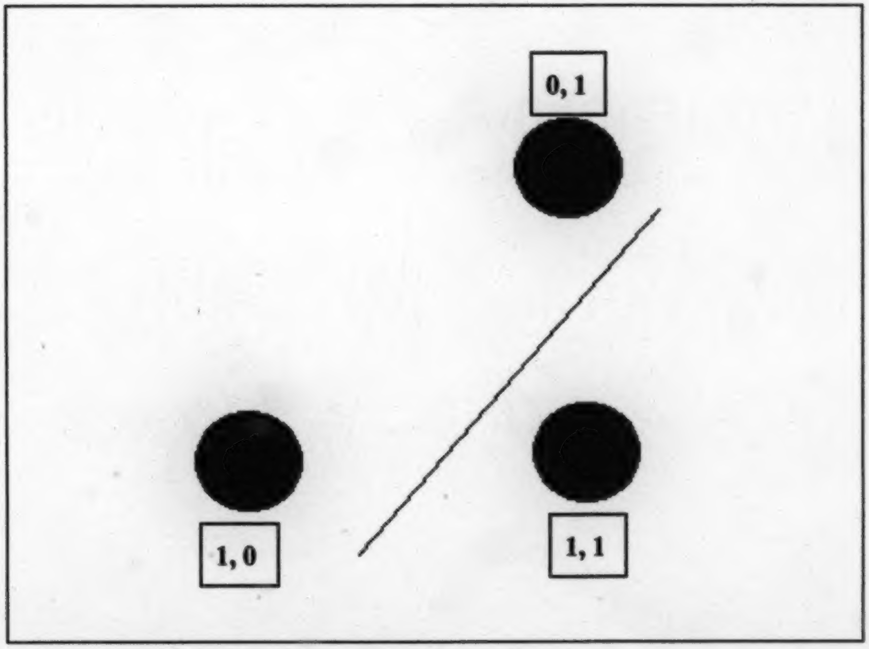

Figure 6-2. Training sets of Delamater et al. (1999), graphically represented. Sets in all conditions require the subjects to discriminate the elements from the compound. Given the number of inputs, this requires one discrimination which is represented by the line.

linearly separable problems only require one discrimination. The positive patterning problem in this study is not exactly a true logical AND problem; likewise, the negative patterning problem is not a typical logical X-OR.

Delamater *et al.* (1999) use a multilayer network in their study, modeled after Rumelhart, *et al.* (1986). Multilayer networks are required for solving linearly non-separable problems. Recall, as discussed in Chapter 3, Minsky and Papert's (1969/88) criticism of perceptron networks included the proof that these networks were incapable of providing a solution to a linearly non-separable problem. This criticism brought about the need for multi-layer models. When a task is linearly separable, as the negative and positive patterning tasks are in the study described above, a perceptron is powerful enough to provide a solution. In theory, the problems in the Delamater *et al.* study can be solved with a simple perceptron.

This study explores this hypothesis and the notion that this simplest of connectionist architectures may provide an interesting and plausible solution to negative and positive patterning when networks are pretrained on selected elements of the training set where the more powerful connectionist network failed to do so. The behaviour that is required by an appropriate model of this type of discrimination learning is that pretraining of elements facilitates learning in both negative and positive patterning.

Method

Networks

There were 25 networks in each condition. Each had 6 inputs and an output unit (see Figure 6.3). The output unit of all networks was influenced by a *bias*. A bias is an

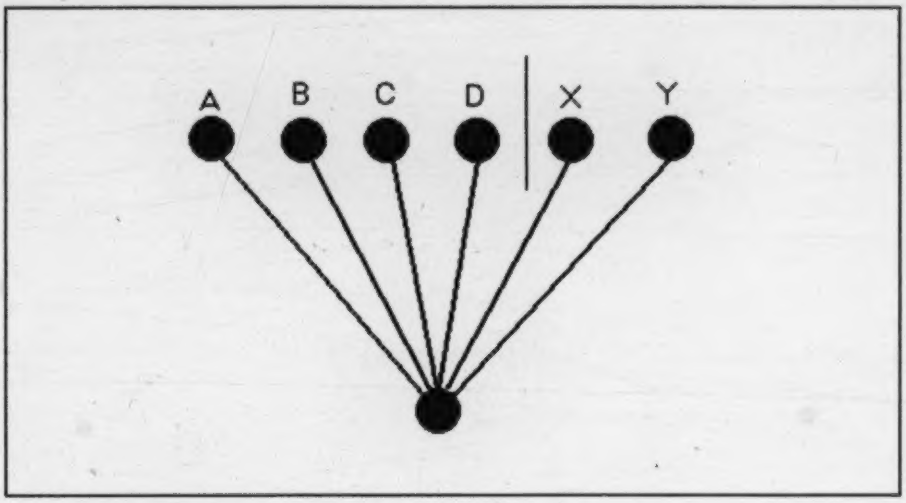

Figure 6-3. Perceptron of Study 6.

extra source of activation that influences only the unit to which it is attached. Although a bias is not technically a processing unit, it is often useful to think of the bias as a unit (Bechtel & Abrahamsen, 1991) that maintains a constant activation of 1 across all input patterns. The bias comes from the weight between the "bias unit" and the output unit. While the activation of the "bias unit" remains fixed, the weight is modifiable and *biases* the output unit. The input units were binary units and were either off (value of 0) or on (value of 1). The activation function of the output unit was a step function. All connection weights were randomly started between -.5 and +.5.

Input Coding

The input patterns were the same as in the Delamater *et al.* (1999) simulation. Cues A, B, C and D corresponded to the first 4 (binary) input units of the network with a value of 1 indicating the presence of a particular cue and a value of 0 indicating its absence. Background cues X and Y were coded in the last two input units in the network and were, again, either off (0) or on (1).

Training

The perceptrons were trained with the Widrow-Hoff learning rule (Widrow & Hoff, 1988). Connection weights were adjusted after each pattern presentation. The learning rate was .1.

Pretraining. Networks in all conditions were pretrained the same way. The pretraining set consisted of the following patterns AX+ (100010+), BX- (010010-), CY+ (001001+), DY- (000101-). A network was considered to have converged, or solved the problem, when its output response was within .1 of the expected response for each of the patterns in the pretraining set. Networks were then required to solve either a positive patterning or negative patterning problem, either with previously reinforced elements (elements A and C) or with previously non-reinforced elements (elements B and D).

Positive and Negative Patterning. In the positive patterning, previously reinforced condition, pretrained networks solved a problem in which elements A and C were used with their corresponding background cues (AX-, CY-, AXCY+). In the positive patterning, previously not reinforced condition, pretrained networks solve a problem using elements B and D with their corresponding background cues (BX-, DY-, BXDY+).

In the negative patterning, previously reinforced condition, pretrained networks solved a negative patterning problem using elements A and C (AX+, CY+, AXCY-). In the previously not reinforced condition, they solved a negative patterning problem using elements B and D (BX+, DY+, BXDY-).

Networks were pretrained and trained until they reached an output state less than .1 from the expected output for each pattern. Number of epochs to convergence was recorded for each network.

Results and Discussion

Pretraining

Cell means and variability for number of epochs to convergence are presented in Table 6-1 for the pretraining phase of the study. When the networks are separated into conditions, there is no significant difference between means in the pretraining phase of the experiment [$F(3,96) = 1.47$; $p > .30$]. As the pretraining phase occurs before there are any differences between the groups, this is as expected. Pretraining specifics are not mentioned in other studies in this chapter. The pretraining phase did not vary in the

		Negative Patterning		Positive Patterning	
		mean	s.d.	mean	s.d.
Pretraining	Previously Reinforced	7.0	2.9	7.2	2.6
	Previously Not Reinforced	5.7	2.9	7.2	3.1
Training	Previously Reinforced	12.6	10.4	10.0	8.7
	Previously Not Reinforced	6.9 (n = 23)	6.1	15.2	12.0

Table 6-1. Means and standard deviations for Study 6.

studies in this chapter and, while pretraining is a manipulation, it is assumed not to vary, either across conditions within studies, or across studies.

Patterning

Cell means and variability for number of epochs to convergence during the training phase are also reported in Table 6-1. There is a significant difference between the positive patterning, previously reinforced versus previously not reinforced group [t pooled (48) = 1.75; $p < .05$, one tailed]. Previous reinforcement of elements facilitates positive patterning learning. For negative patterning, however, although a significant difference is found between the previously reinforced and previously not reinforced groups [t pooled (46) = 2.29; $p < .05$], the difference is not in the expected direction; previous reinforcement clearly does not facilitate the learning of negative patterning in these perceptrons. Two networks were excluded from analysis in the negative patterning, previously not reinforced group as they were identified as outliers (values = 28, 38; see

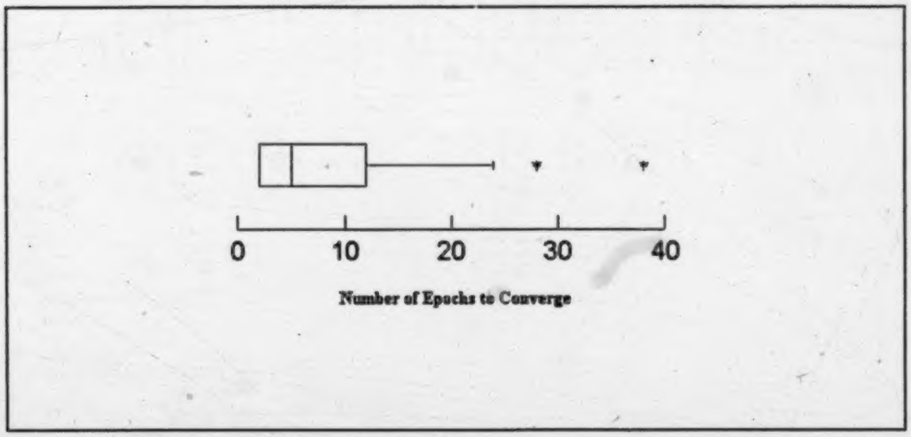

Figure 6-4. Study 6: Perceptron network, negative patterning, previously not reinforced. Box plot identifies 2 outliers.

box plot in Figure 6-4).

This model, then, does not fit the experimental data of Delamater *et al.* (1999). From the discussion earlier, we know that we cannot yet say that connectionist networks do not learn the task as the rats do because a perceptron trained with a Widrow-Hoff training rule is one of a large family of connectionist networks. Rather than simply moving to another network, however, let us return to an assumption made earlier in this study about the training sets used in the simulation. Although the training set used in this study matches the training set used in the simulations in the Delamater *et al.* study, does it really match the training set given to the rat sample?

Consider an appetitive negative patterning scenario in which rats learn to respond to the presentation of element A alone, or element C alone to receive reinforcement, but not to the AC compound stimulus. Is this all the organism learns in this scenario? Actually, the rat also learns that in the absence of element A or element C (for example, during inter-trial intervals), it shouldn't bother to respond, because it is never reinforced in the absence of all stimuli. The networks in this study have never been exposed to a situation in which neither A nor C is presented. To properly simulate the learning environment of the rats in the experimental study, it is necessary to include a "null trial" in which no elements are presented to the network. Perhaps the training in the Delamater *et al.* study does not produce the expected behaviour for negative patterning because the training set does not accurately simulate negative patterning.

If we include a null trial in the above study, the formal problem that is being solved in the negative patterning condition becomes a typical logical X-OR problem. As an X-OR problem, it is no longer linearly separable as was the case in Study 6. This problem is linearly non-separable and, as such, requires the additional power of a multilayer perceptron to solve it.

The following study is an exploration of the effect of adding null trials to the positive and negative patterning training sets in a multilayer perceptron, while continuing to use the pretraining manipulation of the previous study.

Study 7: Negative and Positive Patterning with Null Trials

Method

Networks

There were 25 networks in each condition. Networks in this study had 6 input units, 4 hidden units, and a single output unit. All processing units had logistic activation functions. Prior to pretraining, connection weights were randomly started between -.5 and +.5. The network used in this study is shown in Figure 6-5.

Input Coding

Inputs were again coded across the 6 input units such that A and B were always presented with background unit X and that C and D were always presented with background unit Y. Elements A, B, C, and D corresponded to the first four input units; elements X and Y were coded in the remaining two input units. All input units were binary units, either off (0) or on (1).

Training

The networks were trained using error back-propagation (Rumelhart *et al.*, 1986) in which an error term derived from the difference between the expected output of the

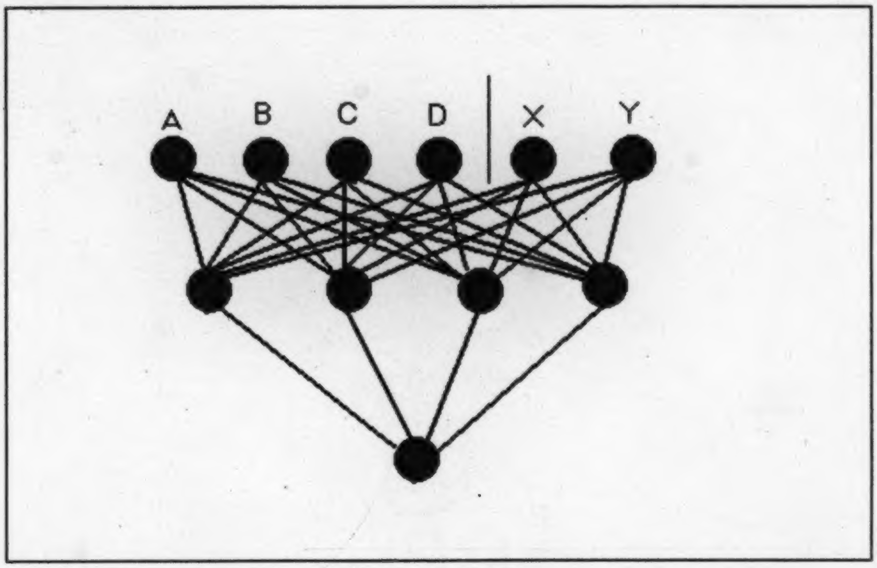

Figure 6-5. Multilayer perceptron used in Studies 7, 8-A, and 8-B.

network and the actual output of the network is propagated back through the network and the connection weights are adjusted. Presentation of the input patterns was randomized within conditions. Connection weights were adjusted after each pattern presentation. The learning rate was 0.1 and all networks were pretrained and trained with a momentum term of .9, in the same way that the networks in the Delamater *et al.* (1999) study were trained.

Pretraining. Networks in all 8 conditions were pretrained in the same way. The pretraining set consisted of the following patterns AX+ (100010+), BX- (010010-), CY+ (001001+), DY- (000101-). A network was considered to have converged, or solved the problem, when its output response was within .1 of the expected response for each of the patterns in the pretraining set. Networks were then either trained on positive patterning or negative patterning, either using the previously reinforced elements or the previously not reinforced elements.

Positive patterning. In the control condition, positive patterning was the same as in Study 6. In all "with nulls" conditions, a null trial (no input units "on", no reinforcement; coded as 000000-) was added to the training set to make the positive and negative patterning problems true to the logical AND and X-OR problems, and more similar to the learning situation that the comparison rat sample faced. In the experimental, previously reinforced condition, pretrained networks were required to solve a positive patterning problem in which cues A and C were used (Null-, AX-, CY-, AXCY+). In the experimental, previously not reinforced condition, networks were required to solve a positive patterning problem in which cues B and D were used (Null-, BX-, DY-, BXDY+). Networks were trained until their output for each pattern fell within .1 of the expected output for the pattern.

Negative patterning. In the control, "without nulls" condition, negative patterning was trained in the same way as in Study 6. In the experimental groups, the null set was added to the training sets. In the experimental, previously reinforced condition, pretrained networks solved a negative patterning problem with elements A and C (Null-, AX+, CY+, AXCY-). The experimental, previously not reinforced condition required networks to solve the negative patterning problem with previously not reinforced elements, B and D (Null-, BX+, DY+, BXDY-). Again, networks were trained until they reached an output state less than .1 away from the expected output for each pattern.

Results and Discussion

Cell means and standard deviations for the number of epochs to convergence for

		Negative Patterning		Positive Patterning	
		mean	s.d.	mean	s.d.
Control (Without Null Trials)	Previously Reinforced	316.3 (n = 24)	47.6	9.6	.7
	Previously Not Reinforced	175.9	7.6	298.7 (n = 22)	29.8
Training With Null Trials	Previously Reinforced	409.9	71.0	169.6	6.6
	Previously Not Reinforced	328.7 (n = 24)	50.9	366.2	58.8

Table 6-2. Cell means and standard deviations for Study 7.

networks in all patterning conditions can be found in Table 6-2. All networks converged: in all conditions, all 25 networks solved the problem.

Given the large differences in the standard deviations of groups that are to be compared, F-tests were run to determine whether the assumption of homogeneity of variances was met for these groups. In the negative patterning paradigm, control group, the variances are significantly different [F (23,24) = 39.2; $p < .01$]. In the negative patterning paradigm, with null trials group, the variances are not significantly different [F (24,23) = 1.36; $p > .01$]. In the positive patterning paradigm, control group, the variances are significantly different [F (21,24) = 1812.3; $p < .01$]. In the positive patterning, with null trials group, the variances are, again, significantly different [F(24,24) = 79.4; $p < .01$]. Non-parametric, two-sample Kolmogorov-Smirnov tests were done for both

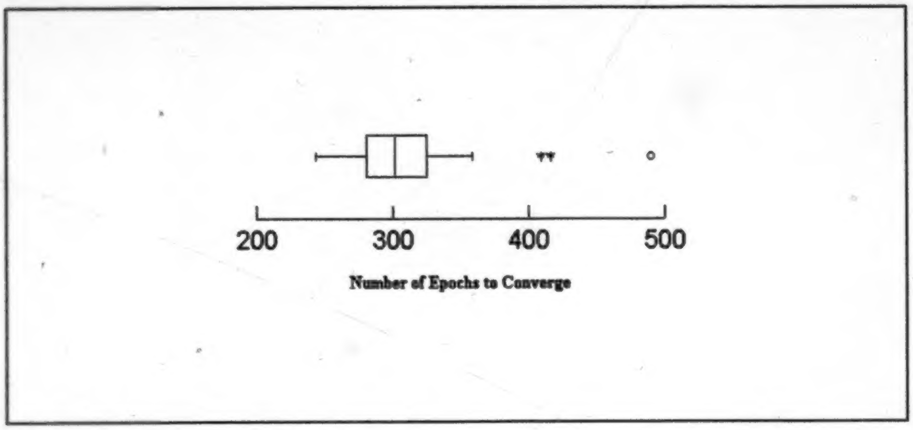

Figure 6-6. Study 7: No nulls, negative patterning, previously not reinforced. Box plot identifies 3 outliers.

conditions in the positive patterning group and for the control group in the negative patterning paradigm. Differences between means were tested in the remaining condition using a *t*-test.

In the control condition, run with no null trials as in the Delamater *et al.* (1999) study and as in Study 6, these multilayer networks display facilitation in the positive patterning condition when the elements have been reinforced, as expected ($D_f = 1.0$; $p < .01$, one-tailed). Three networks were excluded from the previously not reinforced condition as they were identified as outliers (values = 409, 416, 490; see box plot in Figure 6-6). In the negative patterning control condition, however, networks do not display facilitation for reinforced elements, in fact the difference is in the opposite direction and is significant ($D_f = 1.0$; $p < .01$); pretraining inhibits the learning of negative patterning for the networks in this condition. One network was excluded from the previously reinforced condition as it was identified as an outlier (value = 514; see box

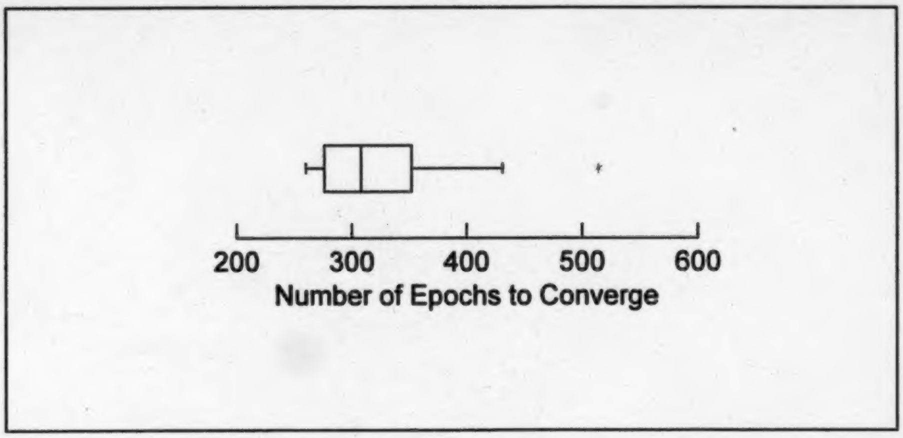

Figure 6-7. Study 7: No nulls, negative patterning, previously reinforced. Box plot identifies 1 outlier.

plot, Figure6-7).

When null trials are added to the training set, networks continue to demonstrate facilitation for the learning of positive patterning ($D_f = 1$; $p < .01$, one-tailed). One of the networks was excluded from analysis in the positive patterning, previously not reinforced condition, as it was a clear outlier (value = 503; see box plot in Figure 6-8). In the negative patterning paradigm, the relationship is, again, not in the expected direction [*t* pooled (47) = 4.6; $p > .01$, one tailed]; in this condition, pretraining clearly does not facilitate learning negative patterning.

This multilayer network with null trials included in the training set still does not fit the experimental data in the negative patterning condition, as there is no facilitation demonstrated in the condition in which previously reinforced elements are used. Given that the model still does not fit the data, there must be something missing in the model. The move taken by Delamater *et al.* (1999) at this point was to add direct connections

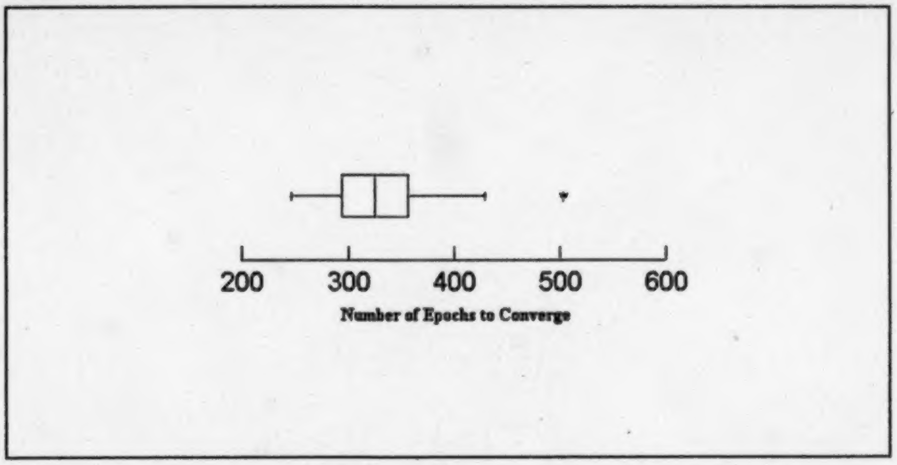

Figure 6-8. Study 7: With nulls, positive patterning, previously not reinforced. Box plot identifies 1 outlier.

from input layer to output layer. Since the model contains sufficient power to solve the problems that it is required to solve, rather than boosting the power by manipulating the architecture, the following study is concerned with *learning* in the networks, and whether the type of learning done by networks in Study 7 can be appropriately compared to the learning situation of the comparison group.

Study 8-A: Should Networks be Losing Momentum?

As mentioned earlier, there are as many varieties of connectionist models as there are researchers that use them, and evaluating the plausibility of these models is important.

A model of learning and its learning environment are made up of a large number of components that influence learning such as the network architecture and the structure of the training set. Not all of these components are justified by the theory they are meant to model. The discrepancy between the components of the system and the theory that the system simulates is the reason for the gap that exists between the "loose level of verbal theorizing and the tight level of description required for a program" (Lewandowski, 1993, p. 240). This gap is bridged by a number of decisions that must be made by the modeler. Those who build and evaluate the plausibility of such models, particularly of learning, need to differentiate between *learning-relevant* components that are part of the theory that is modeled, and *engineering-relevant* components that are included in connectionist models only as part of a design decision. Learning-relevant aspects of connectionist systems are "fair game" for testing; a feature such as distributed representations that is central to a theory that a network models, ought to be explored for plausibility. Not all connectionists have been interested in learning and have introduced features to their models that are interesting from an engineering or a statistical perspective – by making networks converge more quickly, or with less units, or in a manner more easily computed (Dawson & Shamanski, 1994).

Some of these engineering-relevant components have become standard features of automated PDP packages and are incorporated into networks that are then used as psychological models. Although these components may not be intentionally included in the theory of learning that is being evaluated, they indeed are part of the model that is actually being *tested*. As part of the strategy of evaluating networks, these superfluous components should be kept to a minimum. One such feature of connectionist networks is *momentum*, that is described in more detail below.

Connectionist Learning And Momentum

PDP networks learn through a process of adjusting the value of the weighted connections between processing units. For example, consider one popular learning rule called backpropagation of error (Rumelhart, *et al.*, 1986). In backpropagation, a difference between actual network output and expected network output is computed and the error is propagated back through the network. The network then adjusts its weights on the basis of this error: when the error is large, the weight adjustment is large and much learning occurs; when the error is very small, there is little or no change in the connection weights.

In backpropagation, network weights at time $t + 1$ are adjusted according to the following equation:

$$\Delta_p w_{ji(t+1)} = \eta \delta_p o_{pi}$$

in which η is a constant that affects the rate of learning, δ_p is the error associated with the difference between the output of the network and the expected output of the network upon presentation of pattern p, o_{pi} is the i^{th} element of the output pattern associated with the input pattern p, and $\Delta_p w_{ji}$ is the change of the connection weight between the i^{th} and the j^{th} unit when pattern p is presented to the network at time $t + 1$.

Momentum is a component of the backpropagation learning rule that causes the network to adjust the connection weights in a similar direction to that in which they have previously been changed. When momentum is used in a network, one calculates the weight change at time $t + 1$ according to the equation described above. Then, one adds this weight change to the previous weight change (i.e., the weight change at time t) scaled by a constant. This constant is the momentum term. In other words, when momentum is used, learning is governed by the following equation:

$$\Delta_p w_{ji(t+1)} = \eta \delta_p o_{pi} + \alpha \Delta_p w_{ji(t)}$$

in which α is the momentum term.

The advantage of momentum is that it prevents the network from oscillating between two alternatives that would otherwise keep the network from moving toward a solution. Momentum is useful from an engineering perspective – it allows networks to converge faster – however, it is not a feature of a connectionist network that is relevant for a psychological theory. In particular, we know of no evidence that momentum governs associative learning in animals or humans.

In the example from the discrimination learning literature cited above, Delamater *et al.* (1999) used a connectionist network with a relatively small learning rate ($\eta = 0.1$) and used a large momentum term ($\alpha = 0.9$). This study focuses on the effect that this design decision may have had on the behavior that their model produced and explores the possibility that manipulating this variable may produce more plausible learning in multilayer networks.

Method

Networks

Network architecture was the same as in the previous study. Networks had 6 input units, 4 hidden units and one output unit. Processing units had logistic activation functions. Input units were fully connected only to units in the hidden layer; hidden units were connected to the output unit. Prior to training, connection weights were randomly started between -.5 and +.5. Networks that did not train were excluded from the study. There were 25 networks in each condition.

Input Coding

Input patterns were coded in the input layer, again so that elements A and B were always presented with background stimulus X and that elements C and D were always presented with background unit Y. Input units 1 to 4 corresponded to elements A, B, C, and D. Input units 5 and 6 corresponded to the background units. All input units were binary units, either off (0) or on (1).

Training

Two sets of input patterns were used. In one condition, null patterns were not included. This group was included as a control, to determine whether the effect of removing the momentum term was independent of the effect of including null trials in the training sets. In the second condition, null patterns were included.

Networks were pretrained in the same way as in previous experiments in this chapter. They were then required to solve either a negative or a positive patterning problem, as before, using either previously reinforced elements or non-reinforced elements, either with or without a null pattern in the training set. All networks were trained without momentum.

Positive Patterning. In the conditions with no null trials, there were 3 input patterns in each set. In the previously reinforced condition, networks solved a positive patterning problem using cues A and C (AX-, CY-, AXCY+). In the previously not reinforced condition, networks solved a problem using elements B and D (BX-, DY-, BXDY+).

In conditions with null trials included, there were 4 patterns in each set. In the previously reinforced condition, networks were trained on input sets using cues A and C (Null-, AX-, CY, AXCY+). In the previously not reinforced condition, networks were trained on patterns using non-reinforced cues, B and D (Null-, BX-, DY, BXDY+).

Negative Patterning. In the conditions with no nulls, the training set in the previously reinforced condition consisted of patterns using cues A and C (AX+, CY+, AXCY-). In the previously not reinforced condition, training sets included patterns using cues B and D (BX+, DY+, BXDY-).

In the conditions in which null patterns are included, training sets in the previously reinforced condition are (Null-, AX+, CY+, AXCY-) and in the previously not reinforced condition are (Null-, BX+, DY+, BXDY-).

All networks in all conditions were trained until their output fell within .1 of the expected output for each pattern in the training set. Number of epochs to convergence was recorded for each network.

Results and Discussion

The eight cell means and standard deviations for number of epochs to convergence are reported in Table 6-3. As the standard deviations appear different from one another, F tests were done to determine whether they meet the homogeneity of variance constraint within conditions. In the positive patterning situation, without null

		Negative Patterning		Positive Patterning	
		mean	s.d.	mean	s.d.
Without Null Trials	Previously Reinforced	2306.0	353.2	106.8	16.9
	Previously Not Reinforced	1722.4	67.2	3061.5 (n = 24)	449.5
With Null Trials	Previously Reinforced	3672.5 (n = 24)	376.6	1717.4 (n = 23)	69.7
	Previously Not Reinforced	3892.4	1076.4	3448.7	537.4

Table 6-3. Means and variances for Study 8-A.

trials, there is a significant difference between variances [F (23,24) = 707.4; *p* < .01]. In the positive patterning situation, with null trials there is also a significant difference between variances [F (22,24) = 59.4; *p* < .01]. In negative patterning, when nulls are not included there is a significant difference between variances [F (24,25) = 27.6; *p* < .01]. When nulls are included, there is also a significant difference between variances [F (24,23) = 8.1; *p* < .01]. All analyses of differences between means in this study will be done using the Kolmogorov-Smirnov non-parametric test.

Facilitation in the positive patterning situation for previously reinforced elements is present in the "no nulls" condition ($D_r = 1.0$; $p < .01$, one tailed). One network was excluded from analysis in the previously not reinforced condition (value = 4988; see box plot in Figure 6-9). Facilitation is also present in the "with nulls" condition ($D_r = 1.0$; $p <$

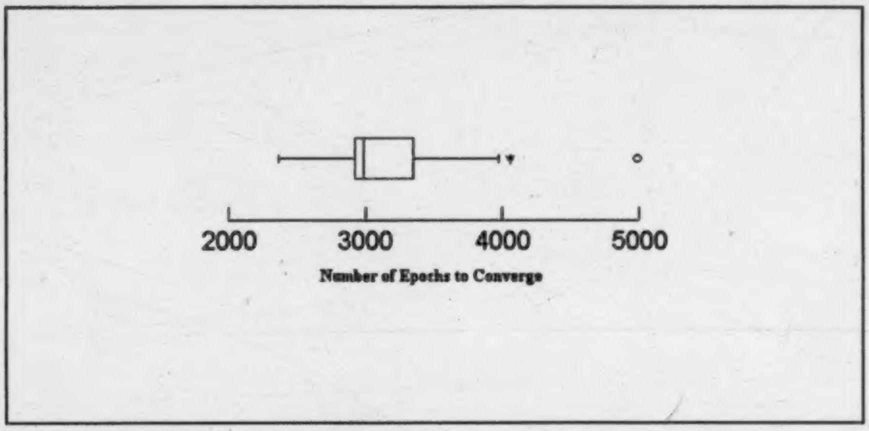

Figure 6-9. Study 8-A: No nulls, positive patterning, previously not reinforced. Box plot identifies one outlier.

.01, one tailed). Two networks were excluded from the previously not reinforced condition as they were identified as outliers (values = 2413, 2497; see box plot in Figure 6-10).

Facilitation in the positive patterning condition when the elements have been

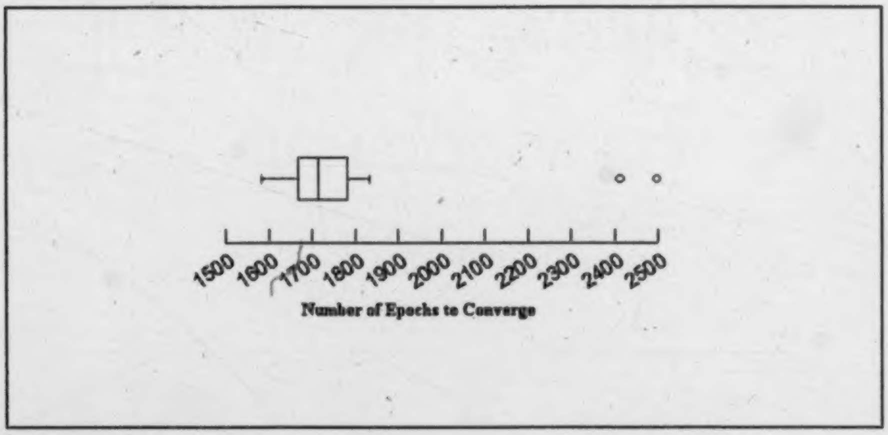

Figure 6-10. Study 8-A: Without nulls, negative patterning, previously not reinforced. Box plot identifies 2 outliers.

previously reinforced appears to be a robust phenomenon as it has been seen in all studies presented in this chapter. It is observed when the network of interest is a perceptron or a multilayer perceptron. It is seen with and without null input patterns in the training set, with and without momentum in the multilayer perceptrons in the studies in this chapter. The critical condition in this set of studies is the negative patterning situation in which facilitation for previously reinforced elements should be observed in the behaviour of networks.

In the negative patterning situation, previous reinforcement of the elements does not facilitate learning without momentum when there are no null patterns in the training set. In fact there is a significant relationship in the opposite direction: pretraining inhibits learning in this condition ($D_f = 1.0$; $p < .01$). One outlier was identified and excluded in the previously not reinforced condition (value = 4021; see box plot in Figure 6-11).

In the negative patterning situation when null patterns are included in the training

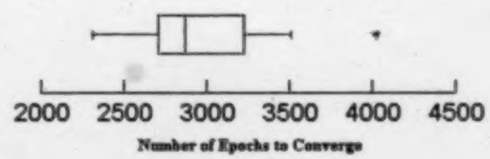

Figure 6-11. Study 8-A. Without nulls, negative patterning, previously not reinforced. Box plot identifies 1 outlier.

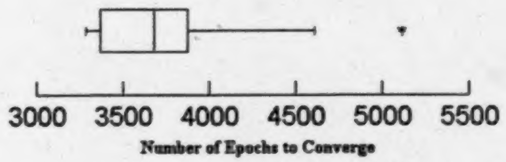

Figure 6-12. Study 8-A: With nulls, negative patterning, previously not reinforced. Box plot identifies 1 outlier.

set and there is no momentum term included in the learning rule, however there *is* a significant difference between the group trained on pretrained elements and on not pretrained elements ($D_f = .44$; $p < .01$, one-tailed). One outlier in the previously not reinforced condition was excluded from analysis (value = 5112; see box plot in Figure 6-12). There is significant facilitation for networks trained on negative patterning using pretrained elements when there are null trials included in the training sets and the momentum term is removed from the learning rule.

In this chapter, we have been concerned with rate of acquisition of a task that has been measured by number of epochs to convergence. This is, of course, only one way to measure learning. In Chapter 4, there was an emphasis on how a task was acquired. In the next study, I present data on how the task was acquired over time, for the networks learning the negative patterning task.

Study 8-B: Momentum and the Course of Learning

One of the concerns that Delamater *et al.* (1999) conveyed about the standard connectionist model that they used was its inability to produce behaviour that looked anything like the behaviour of the rat sample. Rather than observing the rate of acquisition of patterning in a group of networks, in this study, the acquisition of the task is observed. Networks trained, without the use of momentum, on negative patterning with a null trial are compared to the simulations of Delamater *et al.* and to the experimental sample of Delamater *et al.* in terms of how the task is acquired.

Method

The network architecture is the same as in the previous study. The inputs were coded in the same way as in Study 8-A. There were 10 networks in each condition in this study. Network number was limited because of the quantity of data produced by "wiretapping" a network during training.

Training was the same as in Study 8-A for networks learning the negative patterning task. Networks were pretrained as before, then required to solve a negative patterning problem using either previously reinforced elements (Null-, AX+, CY+, AXCY-) or previously not reinforced elements (Null-, BX+, DY+, BXDY-).

In this study, output unit activation was used as a measure of learning. Output unit activation can be interpreted as the *strength* of a response. It should approach 1 for the elements, and 0 for the null trial and the compound in the negative patterning task as training proceeds. Output unit activation was recorded for the presentation of each pattern during the training of these networks.

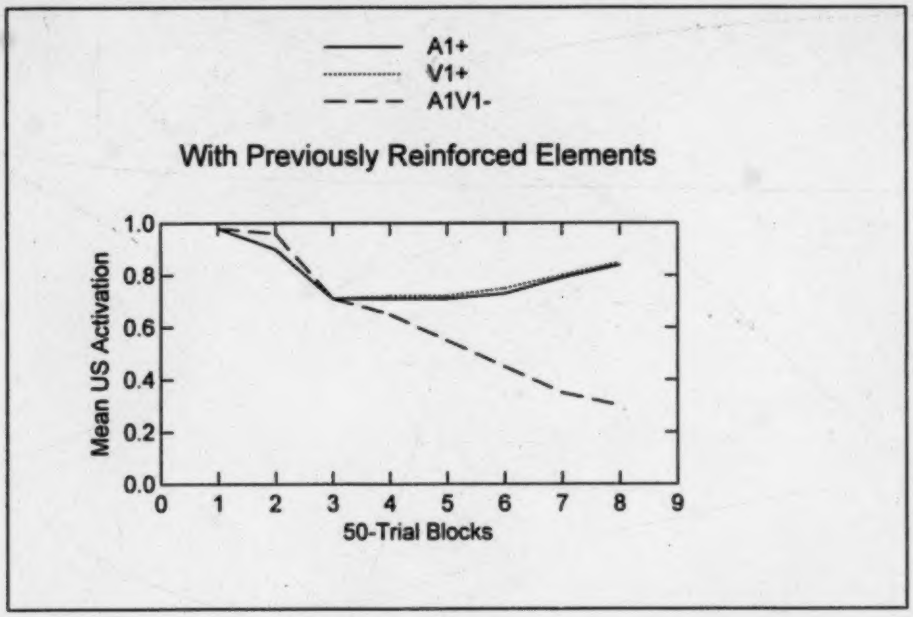

Figure 6-13. Data reproduced from Delamater *et al.* (1999). Negative patterning acquisition for a group of networks learning with momentum, without null trials using previously reinforced elements.

Results and Discussion

The course of learning of the negative patterning task for networks using previously reinforced elements and for networks using previously not reinforced elements is presented in Figures 6-13 to 6-18. Data are presented that were produced by the connectionist model of Delamater *et al.* (1999) in Figures 6-13 and 6-16. The data produced by the rat sample of Delamater *et al.* (1999) are presented in Figures 6-14 and 6-17. Data for the connectionist model presented in this chapter trained with null trials and no momentum term in the learning rule are presented in Figures 6-15 and 6-18[3].

---

[3] Note that the null patterns have been excluded from the graphs to make comparisons across the groups easier. Null patterns were responded to at a low rate in both conditions.

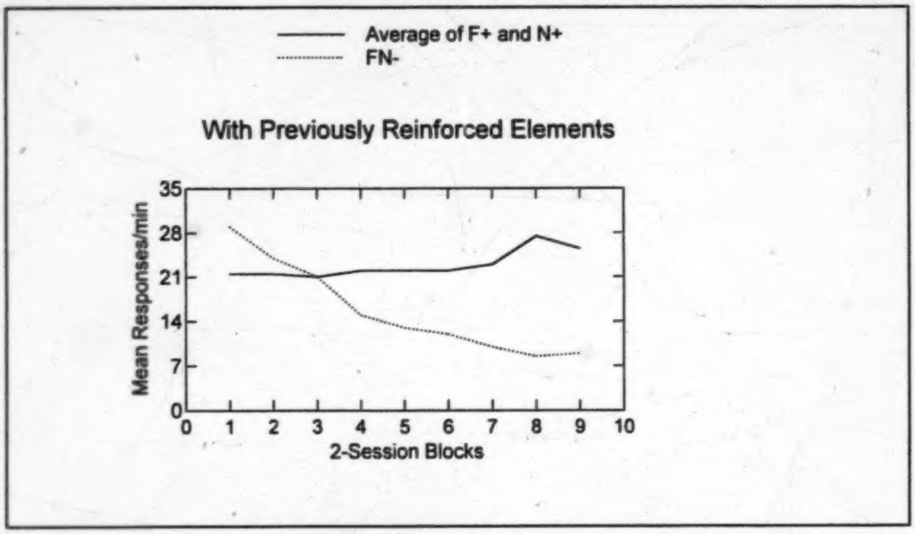

Figure 6-14. Data reproduced from Delamater *et al.* (1999). Negative patterning acquisition for a group of rats using previously reinforced elements.

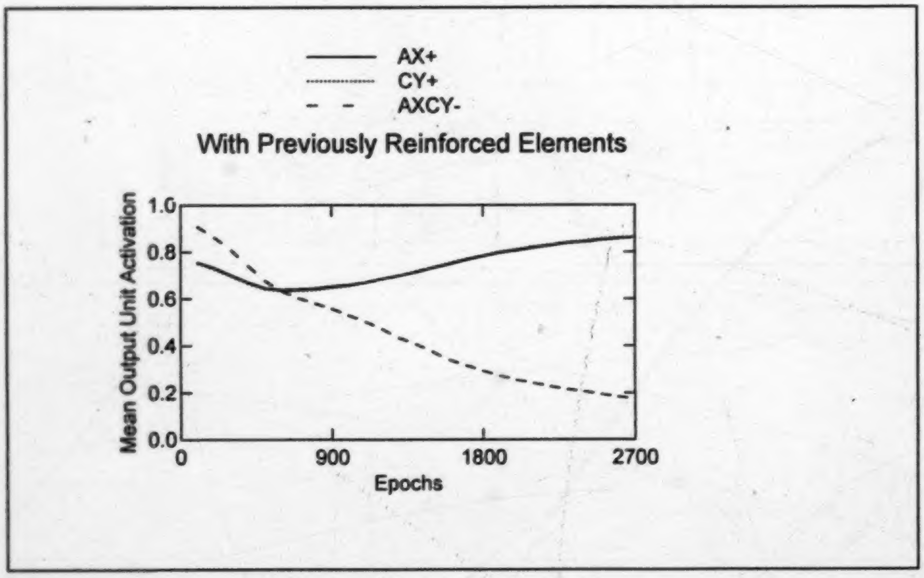

Figure 6-15. Negative patterning acquisition for a group of networks learning without momentum, with null trials using previously reinforced elements.

In the previously reinforced conditions, one can see that the Delamater simulation (Figure 6-13) treats all three trial types (A1+, V1+ and A1V1-) as the same until approximately block 3 (epoch 150) when it begins to acquire the discrimination. In both the rat study (Figure 6-14) and the current simulation without momentum (Figure 6-15), the elements seem to be treated separately from the compound from the start. Both groups display initial excitatory summation: there is a stronger response to the compound that has never been presented, than to either of the components that have been previously reinforced.

In the previously not reinforced conditions, it can be seen in Figure 6-16 that the connectionist model of Delamater and colleagues starts the session with a strong

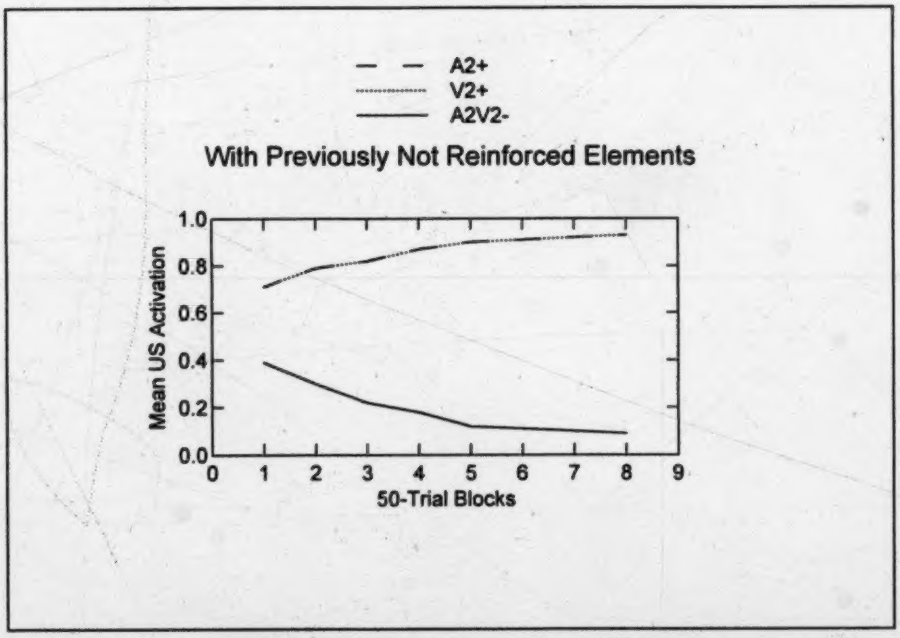

Figure 6-16. Data reproduced from Delamater *et al.* (1999). Negative patterning acquisition for a group of networks learning with momentum, without null trials using previously *not* reinforced elements.

discrimination between the elements and the compound, and continues to improve its performance until the problem is solved. From a neural networks perspective, this is a good solution in that the network solves the problem cleanly and quickly. From an associative learning perspective, however, it is less interesting, as it clearly fails to predict the empirical data in Figure 6-17. The rats display initial excitatory summation in phase I of the acquisition (approximately blocks 1-3): initially, they respond more to the compound than to the elements. In phase II (approximately blocks 3-5) the rats seem to "unlearn" their initial response and by phase III (block 6 and up) they have acquired the

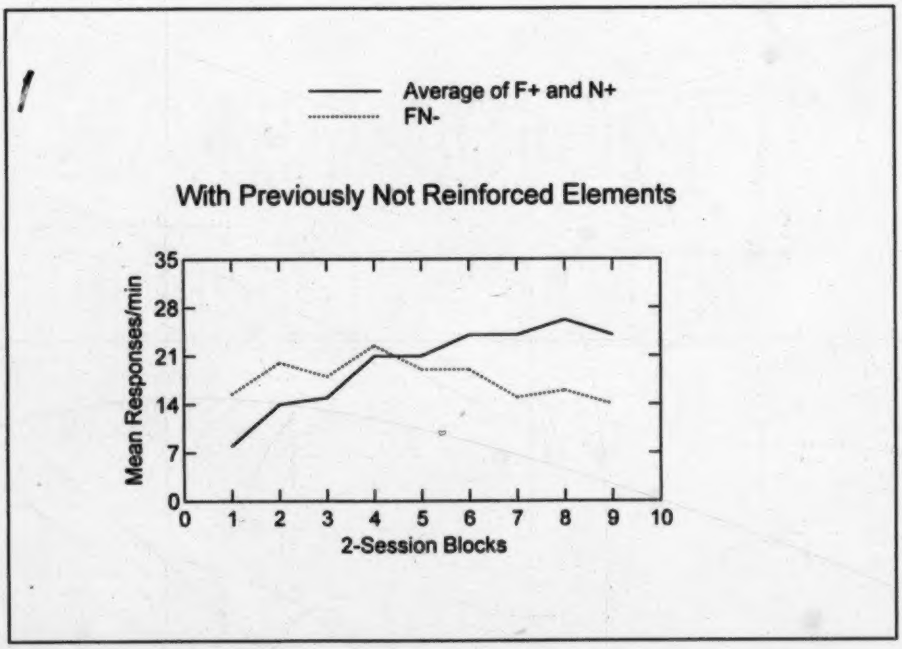

Figure 6-17. Data reproduced from Delamater *et al.* (1999). Negative patterning acquisition for a group of rats using previously *not* reinforced elements.

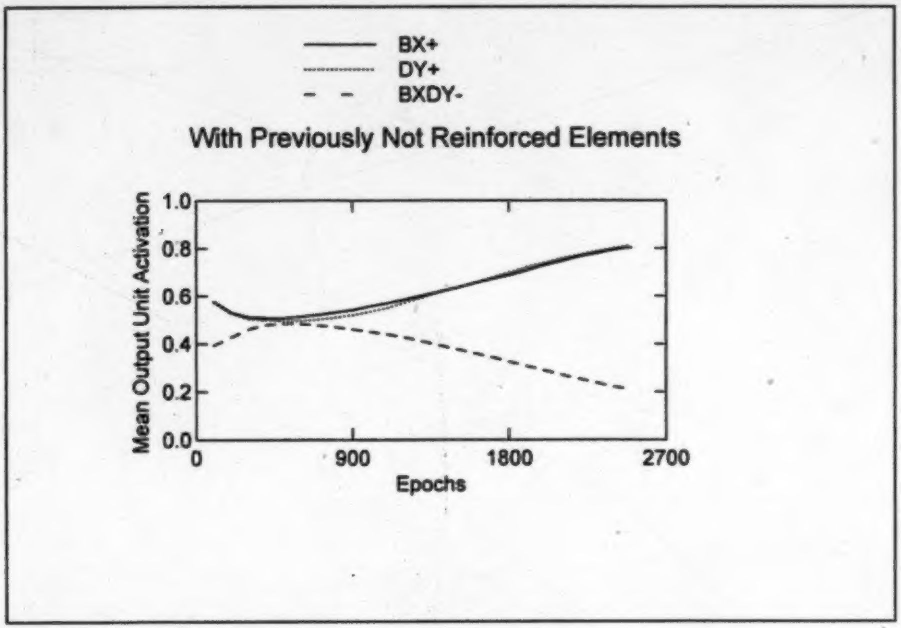

Figure 6-18. Negative patterning acquisition for a group of networks learning without momentum, with null trials using previously *not* reinforced elements.

discrimination. In the no momentum, with nulls condition of the present study, there is no excitatory summation in the previously not reinforced condition (Figure 6-18); rather, the network initially makes an appropriate discrimination. However, the networks seem to unlearn the task in what could be called phase II (from 300-700 epochs), then begin to re-acquire the discrimination at around 800 epochs.

From these results there are several things to note. The first is that the no-momentum model provides results that are similar to the results of the empirical study of Delamater *et al.* in the acquisition of the previously reinforced condition of the negative patterning task. However, the current model without momentum and with null trials is not a perfect fit to the empirical data, as it shows no excitatory summation in the previously not reinforced condition in the negative patterning situation. This is a failure of this iteration of the connectionist networks. However, this model appears to be a better fit to the data than is the PDP model (with momentum) of Delamater *et al.* (1999).

A more significant finding in this study is that, despite their architectural near-equivalence, the PDP model of Delamater *et al.* (1999) and the PDP model presented in this paper show very little resemblance to one another in terms of the course of acquisition of negative patterning. The difference between the two models is the use of momentum and the addition of null trials to the training set. From this we may learn that, in some cases, minor variations in connectionist models and in the training of these models can cause large differences in outcome – differences that may have a large impact on the theories that these simulations are often used to clarify.

General Discussion

In Study 6, we considered the possibility that the problem being posed to the network could be solved with a much simpler system. Positive patterning and negative patterning were solved by a simple perceptron in this study, but, although facilitation was seen in the positive patterning task for networks using pretrained elements, networks trained on the negative patterning task did not show this pattern.

Study 7 began with a question about what organisms really learn when they learn negative and positive patterning. A null trial was added to the training sets to simulate the information that is conveyed to an organism in the absence of any relevant stimuli. With the addition of these null trials, it was necessary to use a multi-layer perceptron. Again, the networks learning the positive patterning task demonstrated facilitation when the elements used for training had been previously reinforced. In the negative patterning situation, however, networks did not demonstrate the facilitation effect as the rats had in the study of Delamater *et al.* (1999).

Study 8-A began with a discussion about what other aspects of the training situation might be responsible for differences between the rat sample and the network sample used in Study 7. A distinction was made between learning-relevant and learning-irrelevant aspects of simulation work. Networks were then trained on negative or positive patterning with pretraining, as before, without a momentum term in the learning rule. Networks learning positive patterning demonstrated facilitation. In the negative patterning condition, networks also demonstrated facilitation, consistent with experimental data.

Study 8-B demonstrated that the course of learning for networks trained without momentum on patterning tasks including a null trial was a good fit to the experimental data of Delamater *et al.* (1999) for the positive patterning task. In the negative patterning condition, the similarities are not as clear between the no-momentum, with nulls network group and the rat sample but the relationship is closer between these groups than between either group and the simulation data of Delamater and colleagues, in which momentum is used and no null trials are included in the training set.

It is important to recall that each connectionist network that is evaluated by a researcher on the basis of its fit to a body of data is one of a large family of connectionist networks that could be considered. The behaviour of different networks in this family of models is different. Therefore, one cannot evaluate the behaviour of a connectionist network on the basis of its failure to fit the behaviour of an empirical sample, and conclude that the sample does not learn in a manner consistent with connectionism.

Evaluating the strength of connectionist models is a necessary pursuit. In this case, we are especially concerned with connectionist models of associative learning. A strategy for testing these models has been pursued in this chapter. A fit to the behaviour of an empirical sample was sought and, indeed, the fit improved with an exploration of the effect of changes to the network architecture. We have examined the effect of altering the architecture and the problem definition and of distinguishing between learning-relevant and learning-irrelevant aspects of simulation. A strategy for *developing* connectionist models is implicit in this process. It is important to consider carefully the implications of various design decisions on the outcome of a study, and to make the simulation situation as close to the experimental learning situation as is possible.

Chapter 7

GENERAL DISCUSSION

In this thesis, I have explored various models of learning, and uses of connectionist networks in the context of discrimination learning problems. In Chapter 2, I introduced contiguity theory and contingency theory, to set the stage for the argument in Chapter 3 that modern connectionist networks may have associative properties, but that they are not *merely* associative. Contingency theory is most properly represented in a simple perceptron, not in a modern, multilayer connectionist network.

Theoretical consistency was the theme of Chapter 3. I argued for the pursuit of connectionist modeling as a *tool* for understanding principles of associative learning, but also as a *theory* that is an extension of current associative theories. Connectionist networks ought not to be considered to be simple implementations of contingency theory, because neural network models necessarily carry with them algorithmic and computational level baggage.

In Chapter 3 it is further argued that several things ought to be considered when using connectionist models. First, researchers ought to understand the theoretical implications of using networks to make predictions or to fit data in learning research. Sometimes, network models used in research have little in common with the theory they are meant to fit. Second, it is important to constrain networks such that as many of the components as possible are consistent with conventions and rules of connectionist modeling and with known principles of learning. The remainder of the thesis focused on the outcomes of various decisions that are made, either intentionally or not, when simulation work is undertaken in learning research.

Summary of Research

All studies are briefly summarized in Table 7-1. In Chapter 4, two studies are presented that are relevant to the question of overfitting in networks. One of the issues mentioned above is the issue of making models that are consistent with conventions and rules of connectionist modeling. One of the expectations of a model is that it is powerful enough to solve a problem but does not overfit the problem of interest. Considering this, and the scientific value of parsimony in theories, Chapter 4 was primarily concerned with

	Manipulation	**Results**
Study 1: Configural Cues in Patterning	- simultaneous negative and positive patterning; with or without pretraining of elements	- excitatory summation when elements pretrained - not when elements not pretrained*
Study 2: Initial Excitatory Summation	- as above but with a value unit output unit	- excitatory summation with pretraining and without pretraining - network interpretation confirms distributed, configural representations
Study 3-A: Forgetting vs. Savings – Rate of Re-Acquisition	- catastrophic forgetting in networks when the dependent measure is rate of re-learning a task - compare a 3-phase group to a 2-phase control	- 3-phase group: relearning is faster than initial learning - 2-phase group: Phase 2 alone does not facilitate learning of Phase 3 (in fact, it inhibits learning of Phase 3)
Study 3-B: Forgetting or Savings – What Does Re-Acquisition Look Like?	- as above but plots of learning in Phase 1, Phase 2 and Phase 3 are compared for the two groups	- while initial SSE is very high in Phase 3 in the 3-phase condition, the Phase 1 task is relearned quickly. Savings in the network is more obvious than forgetting

Study 4: Separate Elements, Separate Tasks	- considers forgetting when the intervening task uses different elements than are used in Phases 1 and 3	- there are significant savings when the elements do not overlap between the initial task and the intervening task - when reacquisition ratios are compared, there is less forgetting when the elements do not overlap than when they do overlap
Study 5-A: Local and Distributed Representations – Savings	- considers forgetting in networks with either 4 hidden units or 2 hidden units - looking for an effect for locality	- comparison of reacquisition ratios reveals no difference in rate of reacquisition (ratio controls for differences in difficulty)
Study 5-B: Local vs. Distributed Representations – "Goodness"	- as above but using SSE before retraining as dependent measure (as Ratcliff)	- comparison of SSE is significant: networks with 4 hidden units are less prone to forgetting when this measure is used - the relevance of this finding to these discrimination learning tasks is questioned
Study 5-C: Local Representations – Network Interpretation	- as above but looks inside a network in each of the conditions to determine the nature of the representations	- find that the local vs. distributed manipulation is not successful as expected - find that two of the hidden units in the 4 hidden unit condition are not necessary - suggest that the finding in Study 5-B (above) may be the result of overfitting rather than local vs. distributed representations
Study 6: A Perceptron Solution to Negative Patterning	- a perceptron is trained on negative and positive patterning (as modeled by Delamater *et al.*) either with or without pretraining	- the perceptron is easily able to solve these problems - when pretraining or no pretraining conditions are compared, facilitation for pretrained conditions is present in positive patterning (as expected) but is not present in negative patterning (facilitation was expected)

Study 7: Negative and Positive Patterning with Null Trials	- as above but validity of the training sets used in the Delamater *et al.* study (and in Study 6) is questioned - null trials are included in the training sets (and compared to a "without nulls" control) - multi-layer perceptrons used instead of perceptrons	- for positive patterning, facilitation for pretrained networks is present in both without nulls and with nulls conditions (as expected) - for negative patterning, neither group demonstrates facilitation in the pretrained condition (where facilitation was expected)
Study 8-A: Should Networks Be Losing Momentum?	- considers momentum and what happens to learning in a network when momentum is removed - networks without momentum, with and without null trials are tested	- facilitation is present for pretrained elements when networks learn positive patterning, both with and without null trials when there is no momentum - facilitation is not present in the negative patterning condition when elements have been pretrained, when there are no null trials in the training set (where facilitation was expected) - there *is facilitation* for pretrained elements when there is no momentum and there are null patterns included in the training set (as per expectation)
Study 8-B: Momentum and the Course of Learning	- the course of learning of networks without momentum and with null trials in the training set are compared to the simulations and the empirical data reported in the Delamater *et al.* study (1999)	- the data in the no momentum, with nulls conditions are more similar to the experimental data of Delamater than the simulations of Delamater as reported in the 1999 study. Some conditions are a very good fit. - however, no excitatory summation is seen in networks learning negative patterning when the elements have not been previously reinforced

Table 7-1. Summary of Studies 1 - 8-B at a glance.

demonstrating the power of connectionist networks. Study 1 illustrated that the discrimination learning problems, negative and positive patterning, can be solved simultaneously with a relatively simple network. This network was powerful enough for the task it was designed to solve, but produced behaviour that did not conform to expectations based on experimental samples. Study 2 presented a network solution to simultaneous negative and positive patterning that produced behaviour closer to expectations. The take-home message for modelers from these two studies is that overfitting a problem may produce solutions that depend on local input-output regularities, rather than solutions in which a network is permitted to develop its own input-output mapping, that is certainly the greatest strength of connectionist models. In Chapter 4, by constraining the number of hidden units in a model, a theory was allowed to emerge in a network, rather than an existing theory simply being instantiated in a network model. The analysis of the internal structure of the network made some predictions concerning the relevance of elemental and configural processes in models of discrimination learning.

While it is important that researchers adhere to the principles of use of connectionist networks, it is equally important that networks produce behaviour that is psychologically valid. In Chapter 5, the phenomenon of catastrophic forgetting was examined, given that some distributed networks are especially prone to it. It was found in Studies 3-A and 3-B that when networks are evaluated in terms of savings, rather than in terms of error after an intervening task, the forgetting that they experience is far from what could be called *catastrophic*. Considering forgetting from a savings perspective seems more relevant when evaluating forgetting in connectionist models of discrimination learning, given that most tasks relevant to the domain require reacquisition of old tasks or acquisition of similar but new tasks.

In Study 4, a question of the nature of the problem was explored. It was deemed a necessary condition of forgetting in networks that the elements used in the initial task and in the intervening task be the same. That is, when the input units are discrete for the two tasks, very little forgetting is seen in the neural networks used in this study.

Comparisons of networks with different numbers of hidden units in Study 5-A showed that within this domain there does not seem to be an effect on forgetting across these groups when reacquisition of a task is considered. In Study 5-B, however, an error measure was considered and an effect for number of hidden units was present. It was concluded that this effect was small and, perhaps, more relevant to other domains that to that of discrimination learning. In any case, the difference between the groups in Study 5-B could not be attributed to a difference between local and distributed networks, as a network analysis in Study 5-C determined that these networks both contained hidden units that were tuned to respond to features in the input space: both contained partially locally distributed representations.

Chapter 6 provided a suggested framework for testing connectionist models. Given that there is no general connectionist model against which specific hypotheses can be tested, I proposed that various members from the family of connectionist networks be evaluated, depending upon the nature of the task and the assumptions of the specific learning task being considered. A starting point is with a network that is the simplest one that can solve the problem of interest. In the case of the negative and positive patterning tasks that have been used throughout the thesis, when null patterns are excluded these tasks can be solved independently with a simple perceptron. This model was explored in Study 6. As this network produced behaviour that did not fit a rat sample that learned either negative or positive patterning after being pretrained on some of the elements, we again considered the nature of the task they were solving. Null patterns were added in Study 7 to simulate the inter-trial-interval for the animal subjects. A multilayer perceptron was needed to solve this task. These networks solved the task did not display facilitation in the negative patterning, previously not reinforced condition, as the experimental sample had done.

When evaluating connectionist models as behavioural models, it is important to isolate those elements of a network that are learning-relevant and to minimize the influence of aspects of the networks that are only engineering-relevant. In Studies 8-A and 8-B it was demonstrated that, by eliminating the momentum term from the learning rule and maintaining the null patterns in the training set, the networks produced behaviour more similar to experimental models.

Implications for Associative Learning Theory and Connectionism

In this thesis, I have presented data and theory that support the use of connectionist models in theories of associative learning. I have promoted the proposition that connectionist models provide powerful, plausible accounts of learning. Yet I have not presented a particular connectionist model to which I am particularly committed. It is one of the goals of this set of studies to explore proper uses of the *family* of connectionist models in the context of discrimination learning. Connectionist models should be exploited for the power they have and for theory development, rather than as an implementation of theories that are less powerful.

Bechtel and Abrahamsen (1991) claim that "If connectionism can produce plausible, powerful learning mechanisms as well as explanatory models of rule-like behavior, it may take a prominent place in cognitive science as an integration of associationism and cognitivism that has a broader domain of applicability than either of its predecessors" (p. 103). Connectionist models are capable of producing rule-like behaviour, as can be seen in studies involving network interpretation (Berkeley, *et al.*, 1995; Christiansen & Chater, 1992; Dawson, 1998; Dawson, *et al.*, 1997; 2000; under review; Elman, 1990). The question of plausible, powerful learning mechanisms has been the focus of this thesis and has not been fully answered here. Models will be needed that make *specific* predictions about discrimination learning, that are constructed with concern for the implications of various design decisions laid out in this thesis.

Connectionism provides a family of models of learning that are distinct from other associative models. The account of learning provided by connectionist networks is representationally powerful beyond earlier theories of learning and can be exploited for theory construction. In this thesis, I demonstrate that connectionist models of learning can also provide plausible accounts of learning, when modelers are attentive to the accuracy of the simulation. With explanations or the promise of explanations of learning at the implementational level, the algorithmic level and the computational level, Connectionism has the potential to provide a plausible and powerful extension of associative theory and a more complete, tri-level account of learning.

REFERENCES

Allan, L., G. & Jenkins, H. M. (1983). The effect of representations of binary variables on judgement of influence. *Learning and Motivation, 14,* 381-405.

Ans, B., & Rousset, S. (2000). Neural networks with a self-refreshing memory: Knowledge transfer in sequential learning tasks without catastrophic forgetting. *Connection Science, 12,* 1-19.

Ballard, D. (1986). Cortical structures and parallel processing: Structure and function. *The Behavioral and Brain Sciences, 9,* 67-120.

Bechtel, W. (1985). Contemporary connectionism: Are the new parallel distributed processing models cognitive or associationist? *Behaviorism, 13,* 53-61.

Bechtel, W., & Abrahamsen, A. (1991). *Connectionism and the mind: An introduction to parrallel processing in networks.* Cambridge, MA: Blackwell.

Bellingham, W. P., Gillette-Bellingham, K., & Kehoe, E. J. (1985). Summation and configuration in patterning schedules with the rat and rabbit. *Animal Learning and Behavior, 13,* 152-164.

Berkeley, I. S. N., Dawson, M. R. W., Medler, D. A., Schopflocher, D. P., & Hornsby, L. (1995). Density plots of hidden value unit activations reveal interpretable bands. *Connection Science, 7,* 167-186.

Bever, T. G., Fodor, J. A., & Garrett, M. (1968). A formal limitation of associationism. In T. R. Dixon & D. L. Horton (Eds.), *Verbal Behavior and General Behavior Theory* (pp.582-585). Englewood Cliffs, NJ: Prentice Hall.

Bitterman, M. E. (1953). Spence on the problem of patterning. *Psychological Review, 60,* 123-126.

Bolles, R. C. (1975). *Learning theory.* New York: Holt, Rinehart and Winston.

Carpenter, G. A. (1997) Spatial pattern learning, catastrophic forgetting, and optimal rules of synaptic transmission. In D. S. Levine & W. R. Elsberry (Eds.). *Optimality in biological and artificial networks?* (pp. 288-316). Mahwah, NJ: Lawrence Erlbaum Associates.

Carpenter, G. A., & Grossberg, S. (1988). The ART of adaptive pattern recognition by a self-organising neural network. *Computer, 21,* 77-88.

Chapman, G. B., & Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Memory and Cognition, 18,* 537-45.

Christiansen, M. H., & Chater, N. (1992). Philosophical issues in connectionist modeling: Connectionism, learning and meaning. (Special issue). *Connection Science, 4,* 227-252.

Churchland, P. S., & Sejnowski, T. J. (1989). Neural representation and neural computation. In L. Nadel, L. Cooper, P. Culicover, & M. Harnish (Eds.). Neural connections, mental computation (pp. 15-44). Cambridge, MA: MIT Press.

Couvillon, P. A., & Bitterman, M. E. (1982). Compound conditioning in honeybees. *Journal of Comparative and Physiological Psychology, 96,* 192-199.

Dawson, M. R. W. (1998). *Understanding Cognitive Science.* Malden, MA: Blackwell Publishers, Inc.

Dawson, M. R. W., Medler, D. A., & Berkeley, I. S. N. (1997). PDP networks can provide models that are not mere implementations of classical theories. *Philosophical Psychology, 10,* 25-40.

Dawson, M. R. W., Medler, D., McCaughan, D. B., Willson, L. R., & Carbonaro, M. (2000). Using extra output learning to insert a symbolic theory into a connectionist network. *Minds and Machines, 10,* 171-201.

Dawson, M. R. W., & Schopflocher, D. P. (1992). Modifying the generalized delta rule to train networks of nonmonotonic processors for pattern classification. *Connection Science, 4,* 19-31.

Dawson, M. R. W., & Shamanski, K. S. (1994). Connectionism, confusion, and cognitive science. *Journal of Intelligent Systems, 4,* 215-262.

Dawson, M. R. W., Willson, L. R., McCaughan, D. B., & Medler, D. (under review). A heuristic stopping rule for the k-means cluster analysis of artificial neural networks. Submitted to *Neural Processing Letters*, 2001.

Delamater, A. R., Sosa, W., & Katz, M. (1999). Elemental and configural processes in patterning discrimination learning. *The Quarterly Journal of Experimental Psychology, 52B,* 97-124.

Ebbinghaus, H. (1885). *Memory: A contribution to experimental psychology.* New York: Dover.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14,* 179-211.

French, R. M. (1992). Semi-distributed representations and catastrophic forgetting in connectionist networks. *Connection Science, 4,* 365-377.

French, R. M. (1997). Pseudo-recurrent connectionist networks: An approach to the 'sensitivity-stability' dilemma. *Connection Science, 9,* 354-381.

Gibbon, J., & Balsam, P. (1981). Spreading association in time. In C. M. Locurto, H. S. Terrace, & J. Gibbon (Eds.), *Autoshaping and conditioning theory* (pp. 219-253). New York: Academic Press.

Gluck, M. A. (1991). Stimulus generalization and representation in adaptive network models of category learning. *Psychological Science, 2,* 50-55.

Gluck, M. A, & Bower, G. H. (1990). Component and pattern information in adaptive networks. *Journal of Experimental Psychology: General, 119,* 105-109.

Gluck, M. A, & Bower, G. H. (1988). Evaluating an adaptive network model of human learning. *Journal of Memory & Language, 27,* 166-195.

Gluck, M. A., & Myers, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus, 3,* 491-516.

Gluck, M. A., & Myers, C. E. (2001). *Gateway to memory: An introduction to neural network modeling of the hippocampus and learning.* Cambridge, MA: MIT Press.

Gluck, M. A., & Thompson, R. F. (1987). Modeling the neural substrates of associative learning and memory: A computational approach. *Psychological Review, 94,* 176-191.

Gorman, R. P., & Sejnowski, T. J. (1988). Learned classification of sonar targets using a massively parallel network. *IEEE Transactions: Acoustics, Speech, and Signal Processing,*36, 1135-1140.

Grossberg, S. (1987). Competitive learning: from interactive activation to adaptive resonance. *Cognitive Science, 11,* 23-63.

Guthrie, E. R. (1935). *The psychology of learning.* New York: Harper.

Hallam, S. C., Grahame, N. J., Harris, K., & Miller, R. R. (1992). Associative structures underlying enhanced negative summation following operational extinction of a Pavlovian inhibitor. *Learning and Motivation, 23,* 76-82.

Hanson, S. J. (1990). Learning and representation: Tensions at the interface. *Behavioral and Brain Sciences, 13,* 511-515.

Hanson, S. J., & Burr, D. J. (1990). What connectionist models learn: Learning and representation in connectionist networks. *Behavioral and Brain Sciences, 13,* 471-518.

Heatherington, P. A., & Seidenberg, M. S. (1989). Is there 'catastrophic interference' in connectionist networks? *Proceedings of the 11th Annual Conference of the Cognitive Science Society,* 26-33.

Hebb, D. O. (1949). *The organization of behaviour.* New York: Wiley.

Hilgard, E. R. (1956). *Theories of learning* (2nd ed.). New York: Appleton-Century-Crofts.

Hinton, G. E. (1989). Connectionist learning systems. *Artificial Intelligence, 40,* 185-234.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart, J. L. McClelland, & the PDP Group (Eds.). *Parallel distributed processing: Vol. 1.* (pp. 77-109). Cambridge, MA: MIT Press.

Hull, C. L. (1943). *Principles of behavior.* New York: Appleton-Century-Crofts.

Jenkins, H. M., Barnes, R. A., & Barrera, F. J. (1981). Why autoshaping depends on trial spacing. In C. M. Locurto, H. S. Terrace, & J. Gibbon (Eds.), *Autoshaping and conditioning theory* (pp. 255-284). New York: Academic Press.

Kehoe, E. J. (1986). Summation and configuration in conditioning of the rabbit's nictitating membrane response to compound stimuli. *Journal of Experimental Psychology: Animal Behavior Processes, 12,* 186-195.

Kehoe, E. J. (1988). A layered network model of associative learning: Learning to learn and configuration. *Psychological Review, 95,* 411-433.

Kehoe, E. J. (1990). Classical conditioning: Fundamental issues for adaptive network models. In M. Gabriel, & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 389-420). Cambridge, MA: The MIT Press.

Kehoe, E. J. (1998). Can the whole be something other than the sum of its parts? In C. D. L. Wynne, & J. E. R. Staddon (Eds.). *Models of action* (pp. 87-126). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Kehoe, E. J., & Gormezano, I. (1980). Configuration and combination laws in conditioning with compound stimuli. *Psychological Bulletin, 87,* 351-378.

Kehoe, E. J., & Graham, P. (1988). Summation and configuration: Stimulus compounding and negative patterning in the rabbit. *Journal of Experimental Psychology: Animal Behavior Processes, 14,* 320-333.

Kehoe, E. J., Horne, A. J., Horne, P. S., & Macrae, M. (1994). Summation and configuration between and within sensory modalities in classical conditioning of the rabbit. *Animal Learning and Behavior, 22,* 19-26.

Kendler, H. H. (1989). Theoretical controversies in behavioural psychology: Are they resolvable? In Keats, J. A., Taft, R. A. Heath, S. H. Lovibond (Eds.), *Mathematical and theoretical systems. Proceedings of the 24th International Congress of Psychology of the International Union of Psychological Science: Vol. 4.* (pp. 259-267). Amsterdam: North-Holland.

Kuhn, T. S. (1970). *The structure of scientific revolutions.* (2nd ed., enlarged). Chicago: University of Chicago Press.

Lachman, R., Lachman, J., & Butterfield, E. (1979). *Cognitive psychology and information processing.* Hillsdale, NJ: Erlbaum.

Leahey, T. H. (1997). *A history of psychology: Main currents in psychological thought* (4th ed.). Upper Saddle River, NJ: Prentice Hall, Inc.

Lewandowsky, S. (1991). Gradual unlearning and catastrophic interference: A comparison of distributed architectures. In W. E. Hockley, & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lewandowsky, S. (1993). The rewards and hazards of computer simulations. *Psychological Science, 4,* 236-243.

Lewandowsky, S., & Li, S.-C. (1995). Catastrophic interference in neural networks: Causes, solutions, and data. In F. N. Dempster, & C. J. Brainerd (Eds.), *Interference and inhibition in cognition* (pp. 329-361). San Diego, CA: Academic Press.

Mackintosh, N. J. (1974). *The psychology of animal learning.* London: Academic

Press.

Maki, W. S. (1990). Toward a unification of conditioning and cognition in animal learning. *Behavioral and Brain Sciences, 13,* 501-502.

Maki, W. S., & Abunawass, A. M. (1991). A connectionist approach to conditional discriminations: Learning, short-term memory, and attention. In M. L. Commons, S. Grossberg, & J. E. R.Staddon (Eds.). *Neural network models of conditioning and action* (pp. 241-278). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Marr, D. (1982). *Vision.* San Francisco, CA: W. H. Freeman.

McClelland, J. L. (2000). Connectionist models of memory. In E. Tulving, & F. I. M. Craik (Eds.). The Oxford Handbook of Memory (pp. 583-596). New York: Oxford University Press.

McClelland, J. L. (1986). Resource requirements of standard and programmable nets. In D. E. Rumelhart, J. L. McClelland, & the PDP Group (Eds.), *Parallel distributed processing: Vol. 1* (pp. 460-487). Cambridge, MA: MIT Press.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102,* 419-497.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88,* 375-407.

McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological Science, 2,* 387-395.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation: Vol. 23.* New York: Academic Press.

McMillan, C., Mozer, M. C., & Smolensky, P. (1991). The connectionist scientist game: Rule extraction and refinement in a neural network. *Procedings of the Thirteenth Annual Conference of the Cognitive Science Society,* 424-430.

Medin, D. L. (1975). A theory of context in discrimination learning. In G. H. Bower (Ed.), *The psychology of learning and motivation: Vol. 9.* (pp. 263-314). New York: Academic Press.

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 355-368.

Miller, R. R., Barnet, R. C., & Grahame, J. N. (1995). Assessment of the Rescorla-Wagner Model. *Psychological Bulletin, 117,* 363-386.

Miller, R. R., & Matzel, L. D. (1988). The comparator hypothesis: a response rule for the expression of associations. In G. H. Bower (Ed.). *The psychology of learning and motivation: Vol. 22* (pp. 51-92). San Diego, CA: Academic Press.

Minsky, M., & Papert, S. (1988). *Perceptrons* (3rd ed). Cambridge, MA: MIT Press. (Originally published in 1969).

Napier, R. M., Macrae, M., & Kehoe, E. J. (1992). Rapid reacquisition in conditioning of the rabbit's nictitating membrane response. *Journal of Experimental Psychology: Animal Behavior Processes, 18,* 182-192.

Newell, A. (1973). Production systems: Models of control structures. In W. G. Chase (Ed.), *Visual Information Processing* (pp. 463-515). New York: Academic Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Pavel, M. (1990). Learning from learned networks. *Behavioral and Brain Sciences, 13,* 503-504.

Pavlov, I. (1927). *Conditioned reflexes.* London: Oxford University Press.

Pearce, J. M. (1987). A model of stimulus generalization for Pavlovian conditioning. *Psychological Review, 94,* 61-73.

Pearce, J. M. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review, 101,* 587-607.

Pearce, J. M., & Bouton, M. E. (2001). Theories of associative learning in animals. *Annual Review of Psychology, 52,* 111-39.

Peterson, C. R. & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68,* 29-46.

Pylyshyn, Z. W. (1984). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences, 3,* 111-169.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review, 97,* 285-308.

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology, 66,* 1-5.

Rescorla, R. A. (1969). Conditioned inhibition of fear. In W. K. Honig and N. J. Mackintosh (Eds.), *Fundamental issues in associative learning* (pp. 65-89). Halifax: Dalhousie University Press.

Rescorla, R. A. (1992). Hierarchical associative relations in pavlovian conditioning and instrumental training. *Current Directions in Psychological Science, 1,* 66-70.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy, (Eds.), *Classical conditioning II: Current theory & research* (pp. 64-99). New York: Appleton-Century-Crofts.

Robins, A. (1995). Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science, 7,* 123-146.

Robins, A. (1996). Consolidation in neural networks and in the sleeping brain. *Connection Science, 8,* 259-376.

Roiblat, H. L. (1982). The meaning of representation in animal memory. *The Behavioral and Brain Sciences, 5,* 353-406.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65,* 386-408.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP

Group (Eds.), *Parallel distributed processing: Vol. 1* (pp. 318-362). Cambridge, MA: MIT Press.

Rumelhart, D. E. & McClelland, J. L. (1985). Levels Indeed! A response to Broadbent. *Journal of Experimental Psychology: General, 114,* 193-197.

Rumelhart, D. E. & McClelland, J. L. (1986). PDP models and general issues in cognitive science. In D. E. Rumelhart, J. L. McClelland, & the PDP Group (Eds.), *Parallel distributed processing: Vol. 1* (pp. 110-146). Cambridge, MA: MIT Press.

Rumelhart, D. E., Todd, P. M. (1993). Learning and connectionist representations. In D. E. Meyer, & S. Kornblum (Eds.), *Attention and performance 14: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience.* Cambridge, MA: MIT Press.

Sarle, W. S. (1994). Neural networks and statistical models. *Proceedings of the Nineteenth Annual SAS Users Group International Conference,* 1-13.

Schmajuk, N. A. (1997). *Animal learning and cognition: A neural network approach.* New York: Cambridge University Press.

Schmajuk, N. A., & DiCarlo, J. (1992). Stimulus configuration, classical conditioning, and hippocampal function. *Psychological Review, 99,* 268-305.

Schwartz, B., & Robbins, S. J. (1995). *Psychology of learning and behavior (*4th ed.) New York: W. W. Norton.

Seidenberg, M. (1993). Connectionist models and cognitive theory. *Psychological Science, 4,* 228-35.

Seidenberg, M. & McClelland, J. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96,* 523-568.

Shanks, D. R. (1995). *The psychology of associative learning.* Cambridge, GB: Cambridge University Press.

Shanks, D. R. (1991). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 17,* 433-443.

Spence, K. W. (1936). The nature of discrimination learning in animals. *Psychological Review, 43,* 427-449.

Spence, K. W. (1952). The nature of the response in discrimination learning. *Psychological Review, 59,* 89-93.

Sternberg, S. (1970). Mental scanning: mental processes revealed by reaction time experiments. In J. S. Antrobus (Ed.), *Cognition and Affect.* Boston, MA: Little Brown.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88,* 135-170.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning.* Cambridge, MA: MIT Press.

Thorndike, E. L. (1932). *The fundamentals of learning.* New York: Teachers College, Columbia University.

Tolman, E. C. (1932). *Purposive behavior in animals and men.* New York: Century.

Wagner, A. R. (1968). Incidental stimuli and discrimination learning. In G. Gilbert and N. S. Sutherland (Eds.), *Discrimination learning.* London: Academic Press.

Wagner, A. R. (1969a). Stimulus validity and stimulus selection in associative learning. In *Fundamental issues in associative learning* (pp. 90-122). Halifax: Dalhousie University Press.

Wagner, A. R., Logan, F. A., Haberlandt, K., & Price, T. (1968). Stimulus selection in animal discrimination learning. *Journal of Experimental Psychology, 76,* 171-180.

Wasserman, E. A., Elek, S. M., Chatlosh, D. L., & Baker, A. G. (1993). Rating causal relations: The role of probability in judgments of response-outcome contingency. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19,* 174-188.

Wasserman, E. A., & Miller, R. R. (1997). What's elementary about associative learning? *Annual Review of Psychology, 48,* 573-607.

Weidemann, G., & Kehoe, E. J. (1997). Transfer and counterconditioning of conditional control in the rabbit nictitating membrane response. *The Quarterly Journal of Experimental Psychology, 50B,* 295-316.

Widrow, B. & Hoff, M. E. (1988). Adaptive switching circuits. In J. A. Anderson & E. Rosenfeld (Eds.), *Neurocomputing; Foundations of Research* (pp. 126-34). Cambridge, MA.: MIT Press, (Reprinted from 1960 IRE WESCON Convention Record. New York: IRE, pp. 96-104).

Willson, L. R. W., Valsangkar-Smyth, M. A., McCaughan, D. B., & Dawson, M. R. W. (1999, June). Cluster analysis of PDP networks: Two rules for deciding how many clusters to extract. Paper presented at CSBBCS, Edmonton, AB.